UNIVERSITY OF HELSINKI

Exploring Linguistic Diversity in South America

A Case Study of the Tropical Andes and the Peruvian Amazon

Kirsi Kauppinen
Master's Thesis
General Linguistics
University of Helsinki
May 2018

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty	Laitos – Institution – Department	
Humanistinen tiedekunta		
Tekijä – Författare – Author		
Kirsi Kauppinen		
Työn nimi – Arbetets titel – Title		
Exploring Linguistic Diversity in South America: A Case Study of the Tropical Andes and the Peruvian Amazon		
Oppiaine – Läroämne – Subject		
Yleinen kielitiede		
Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä– Sidoantal – Number of pages
Pro gradu -tutkielma	Toukokuu 2018	68

Tiivistelmä – Referat – Abstract

Tutkielman aiheena on kielellisen diversiteetin tutkiminen Etelä-Amerikassa. Tavoitteena on selvittää, miten kielellisiä eroavaisuuksia voidaan mitata, ja miten näitä eroja voidaan selittää. Tarkoituksena on myös kuvata diversiteetin vaihtelua diakronisesti Etelä-Amerikassa ja ehdottaa syitä diversiteetin muutoksille. Yhtäältä tutkielma on siis kvantitatiivinen, mutta toisaalta myös kvalitatiivinen. Tavoitteena on lisäksi vastata seuraaviin kysymyksiin: Missä määrin olemassaolevat sukulaisuussuhteet vaikuttavat kielten välisiin eroihin? Voivatko maantieteelliset ja sosioekonomiset tekijät selittää kielellisiä eroavaisuuksia kielten välillä? Tutkielmassa esitetään myös hypoteesi, jonka mukaan kielten elinympäristö ja kielelliset eroavaisuudet korreloivat keskenään.

Tutkielman teoreettisena taustana toimii kielellisen diversiteetin kuvaaminen, sekä analyysi diversiteetin vaihteluista. Teoriaosassa käsitellään yksityiskohtaisesti eri lähestymistapoja, jotka tukevat tutkielman kvantitatiivista ja kvalitatiivista tutkimusta. Nämä lähestymistavat kuvaavat tarkemmin eri keinoja tutkia kielellisiä eroavaisuuksia, ja ne myös laajentavat kielellisten eroavaisuuksien selittämiseen käytettävää teoriaa.

Tutkimuksen aineisto koostuu yhdeksästä eteläamerikkalaisesta kielestä, joita puhutaan trooppisilla Andeilla Kolumbian ja Ecuadorin alueella sekä Perun Amazonin alueella. Aineisto koostuu kolmesta isolaattikielestä, ja kuudesta eri kielikunnan kielestä. Analyysi suoritetaan vertailemalla kielien kuutta eri rakenteellista piirrettä käyttäen tilastollista menetelmää, joka mittaa kuinka erilaisia kielet oikeasti ovat. Mittaukset perustuvat etäisyysmatriisiin, jossa rakenteelliset piirteet esitetään numeerisina arvoina. Lopputuloksena on kuvaaja, jossa mitatut eroavaisuudet esitetään kaksiulotteisessa tasossa.

Tilastollinen analyysi osoittaa, että kielten eroavaisuuksia voidaan mitata. Tutkielman pohdintaluvussa kuvataan myös syvällisesti, miten maantieteelliset ja sosioekonomiset tekijät voivat selittää kielellisiä eroja. Mitattujen eroavaisuuksien perusteella havaitaan esimerkiksi maantieteellisten sijaintien vaikutus eri kieliin, sillä samalla alueella puhuttavat kielet osoittavat suuria kielellisiä eroja, mikä on selitettävissä kielten elinympäristöjen mahdollistamalla eristyksellä. Lisäksi sosioekonomiset tekijät, kuten kaupungistuminen sekä alkuperäiskansojen sisäryhmäavioliittoisuus, voivat selittää sekä diversiteetin vähenemistä että kielellisten eroavaisuuksien kasvamista. Diversiteetin vaihtelua kuvataan esittämällä myös muita syitä, kuten maanviljelyn kehittyminen ja teollinen vallankumous. Tulosten perusteella voidaan todeta, että kielikuntien viimeiset elossaolevat jäsenet, eli isolaattikielet, osoittavat huomattavia rakenteellisia eroavaisuuksia verrattuina kieliin, joilla on yhä olemassaolevia sukulaisuussuhteita. Tulokset vahvistavat myös hypoteesin, jonka mukaan kielten elinympäristö korreloi kielellisten eroavaisuuksien kanssa. Täten tutkielma tukee käsitystä kielten ja niiden elinympäristön välisestä monimutkaisesta suhteesta.

Avainsanat – Nyckelord – Keywords
kielellinen diversiteetti, kieliekologia, Etelä-Amerikka, tilastollinen analyysi

Säilytyspaikka – Förvaringställe – Where deposited
Keskustakampuksen kirjasto

Muita tietoja – Övriga uppgifter – Additional information
Suom. Etelä-Amerikan kielellisen diversiteetin tutkiminen trooppisilla Andeilla ja Perun Amazonissa.

Table of Contents

List of Figures

List of Tables

1 Introduction

There are approximately 7,000 languages spoken in the world, generating remarkable linguistic diversity (Simons & Fennig 2018). Understanding the diversity among the world's languages is one of the focal points in modern linguistics. Linguists are interested in investigating how languages have actually emerged, and how languages change across space and time. They have understood that languages are not stable systems, but permissive of adapting to their surrounding environments. These variations in diversity have intrigued many academics, such as Dixon (1997) and Nettle (1999). Additionally, several scholars have noted how linguistic diversity is not spread evenly across the globe (Gavin 2014, Nettle 1998, Nichols 1992). This unevenness is quite remarkable, because only 9 % of the world's land area contains approximately 60 % of the world's languages (Nettle & Romaine 2000). What is even more astonishing is the fact how so many different languages are actually spoken within these dense diversity hotspots. There are languages from hundreds of language families and also dozens of language isolates, and yet they all differ from one another despite their close proximities.

One of the diversity hotspots in the world is South America. According to *The Ethnologue*, there are 455 living languages spoken in the area. These languages belong to 108 attested language families, or they are one of the 55 language isolates. Linguistically this diversity is extremely fascinating, because it is not only genetic, but also typological (Adelaar 2004). Additionally, South America is diverse not just in the volume of languages but also in the geographical distribution of the languages. Compared to the rest of the world, the languages spoken in South America are distinctly discontinuous in their distribution around the continent (Dixon & Aikhenvald 1999). This means that some South American languages differ from other languages despite their geographical proximities. Languages can also share similarities, which can be due to the genetic relationship between languages, or as in some cases, the close contact between languages. Considering these aspects it is extremely valuable to study linguistic diversity and to measure linguistic differences in South America, because it can reveal how diverse the languages actually are, and this is exactly what I am doing in this thesis.

Because measuring the linguistic differences of an entire diversity hotspot would be too excessive for this thesis, it suffices to choose a smaller area within South America. I decided to choose

an area encompassing the tropical Andean countries of Colombia and Ecuador and also the Peruvian Amazon. The area is highly diverse not just linguistically, but also in its physical geography. A remarkable feature is the high mountain range of the Andes, in addition to the Amazonian lowlands with their dense river systems. For the study I will include nine languages from the area, six of which represent different language families and three are language isolates. Typologically the language data is not extremely representative, but as a case study this thesis shows how linguistic differences can be researched, and that the study could be done with a much larger data set.

There has been an increase in studying linguistic diversity and the differences between language systems. One of the recent examples is the work of Borin & Saxena (2013), which addresses the various aspects of measuring linguistic differences. However, most of the case studies and methods addressed in their work are only concerned with the genetic classification of languages through the difference measures. Linguistic diversity on the other hand has been studied from several different viewpoints, such as the relationship between biodiversity and linguistic diversity (Gorenflo et al. 2012) and viewing diversity through multilingualism (Edwards 2012). Despite these in-depth investigations, neither the contributions in Borin & Saxena, nor, to my knowledge, any other studies actually measure linguistic differences. This can be achieved by using quantitative methods. In general quantitative methods can be used to compare the typological profiles of languages, which would in turn help understand the degree of linguistic diversity in the world. The basis for my thesis spawns from the extensive implementations these quantitative methods can offer for studying linguistic diversity.

The increase in studies on linguistic differences is due to developments in computational linguistics, which offer more efficient ways of measuring linguistic differences. Understandably, difference is a rather difficult concept to measure. Nonetheless, there are quantitative tools which can render estimates of the differences, and eventually reveal underlying correlations in the data (Hout et al. 2013). A particular statistical method is multidimensional scaling, which enables the spatial representation of statistical measures (Chambers & Trudgill 2004). Multidimensional scaling is used in this thesis in order to measure and visualize the linguistic differences in the language data. The analysis is done by comparing the structural features of languages, which are extracted from the World Atlas of Language Structures, and then analyzed with a statistical computing software R. The result is a map, where similar languages are located in close vicinity to one another and dissimilar languages are located further apart (Hout et al. 2013).

To summarize, the objective of this thesis is to explore linguistic diversity, and especially the linguistic differences between languages spoken in the tropical Andes and in the Peruvian Amazon. I will examine how diverse a set of languages actually are, when the data set consists of language isolates and languages from attested language families which are spoken in a distinct area inside a diversity hotspot. Specifically, I would like to consider the following questions:

- How can linguistic differences be measured? To what extent can existing genealogical relationships affect the linguistic differences between languages?
- What are the possible reasons and explanations for linguistic diversity in South America, and specifically in the tropical Andes and in the Peruvian Amazon?
- Can the results be explained by using an ecological point of view? Can non-linguistic parameters, such as geographical and socioeconomic factors, be used to explain linguistic differences?

In addition to the aforementioned research questions my hypothesis is that there is a correlation between the ecological environment of languages and their linguistic differences.

The aim of this thesis is twofold. First, the aim is to conduct a quantitative study by using a statistical research method, and second, to explain the results of that research qualitatively. The theoretical approach of the thesis is both areal-typological, because I am comparing languages spoken in a specific area, and language-ecological, because I am exploring linguistic diversity and linguistic differences through language ecology. I will also try to explain how and why the overall linguistic diversity in the area has changed diachronically. Essentially this thesis is an interdisciplinary case study which involves a quantitative analysis based on specific languages spoken in a specific area. Hence the answers to the research questions above are meant to be given with respect to the research area and to the language data, and not as universal answers compatible with all languages and all diversity hotspots.

This thesis begins with an overall description of the theoretical framework in Chapter 2, which consists of two distinct parts. First, I will discuss the appropriate theoretical background, which sets the scene for the thesis. Second, I will elaborate on different approaches used to conduct the study. The approaches will give an insight into how linguistic differences can be studied, and they also elaborate on a theoretical level on the linguistic and non-linguistic ways of explaining linguistic differences.

Chapter 3 deals with South America by focusing on the linguistic diversity, physical geography and on the demographics of the continent. In addition, these aspects are discussed in light of the countries situated in the research area. Because the relationship between language and its environment is a complex one, this chapter will aid in understanding the overall background against which the languages studied in this thesis exist.

Chapters 4 and 5 describe the methodological aspects of the thesis. First in Chapter 4, I will thoroughly describe the language data used to conduct the analysis, and go through the process of choosing the linguistic parameters for the study. Second, I will elaborate on the statistical software R, and the actual statistical method, multidimensional scaling, in Chapter 5. In the end I will portray the results of the analysis, which show the multidimensional scaling of the data, i.e. the visualization of the linguistic differences.

In Chapter 6 I will explain why the chosen languages spoken in the tropical Andes and in the Peruvian Amazon are different, and how the area's overall linguistic diversity has changed diachronically. These explanations will underline the relationship between language and its environment, and in the end I will be able to answer my research questions presented in this introductory chapter. Based on the explanations I will also verify the hypothesis. Finally, the concluding remarks are given in Chapter 7.

2 Theoretical framework[1]

In this thesis I will compare a set of languages spoken in the tropical Andes and in the Peruvian Amazon, and then explain the differences between the languages and the linguistic diversity of that area by paying attention to geographical and social environments in which the languages are used. This comparison is done by measuring linguistic differences according to a few attested structural features using a statistical method, which quantifies the differences found between the languages. This study is both quantitative and qualitative. Quantitative statistics is used in order to calculate and visualize the differences in the data set, and the results are explained qualitatively by focusing on linguistic and non-linguistic parameters. With this in mind it is sufficient to include two distinct sections in this chapter for both theoretical background and for theoretical approaches. The theoretical background sheds light on linguistic diversity, specifically on its definition, distribution and fluctuation, and the theoretical approaches give insights on the methods for studying linguistic differences and on the ways I will analyze and explain the results.

So as a whole, this thesis combines two different theoretical approaches. On the one hand, it is areal-typological in its comparison of different languages, and on the other hand it is also language-ecological due to the explanations of the structural differences and the diversity in the area. In this chapter I will first elaborate on the theoretical background, and then I will describe the different approaches.

2.1 Theoretical background

This section focuses on characterizing linguistic diversity, mostly by how it is defined, and how it can change across space and time. The following descriptions will set the overall scene for this thesis.

[1] Even though linguistic diversity is a complex phenomenon, I will not describe either the origins of language or the entire evolution of diversity in this chapter, since both of these topics are extremely wide areas in linguistics. I will also not define the concept of language and what is actually seen as a language. For the sake of this thesis, it is not essential to draw a line between language and dialect or to talk about mutual intelligibility, because what matters in the end is that all the languages spoken in the world have their own systems and structures, and I am only interested in that structural diversity.

2.1.1 Linguistic diversity

The world's languages are related to one another just like biological species are, through several embedded patterns of descent (Gavin et al. 2013). These patterns have generated a spectacular amount of linguistic diversity: according to *The Ethnologue* there are approximately 7,000 languages spoken in the world. Most of these languages have a relatively small number of speakers, and most of them are only used orally. Of these several thousand languages though, the 100 most used ones are spoken by 90 % of the population (Nettle & Romaine 2000). This is definitely a bewildering portion of people. What about the remaining 6,900 or so languages? Those languages are mostly unwritten local community languages and they are found all across the globe. To put things into perspective, as Nettle & Romaine (2000: 32) state, almost "85 % of languages have fewer than 100,000 [speakers]". Despite the unevenness of language speakers across the world, it is clear that the 7,000 languages vary in numerous different ways, since there are areas where dozens or even hundreds of languages are spoken, but they still manage to be linguistically diverse. So across millennia diversifying mechanisms shaped our existence, which resulted in a linguistically diverse world, where languages differ in many ways (Nettle 1999).

Linguistic diversity does not just mean that there are thousands of different languages. Nettle (1999) has divided linguistic diversity into three different subcategories, which are language diversity, phylogenetic diversity and structural diversity. Language diversity refers to the actual number of languages in a given area, which varies quite drastically between countries and continents. Phylogenetic diversity quantifies the number of language lineages in a given area. A general way of measuring phylogenetic diversity is through the concept of language families, which is used by historical and comparative linguists in order to group languages according to their ancestry. *The Ethnologue* lists 145 different language families, excluding creoles, pidgins, isolates, sign languages and unclassified languages. However, high language diversity does not always imply high phylogenetic diversity. As an example, Nettle (1998) describes how there are hundreds of different languages in central Africa, but almost all of them are closely related, belonging to the Bantu language family, which makes the area low in phylogenetic diversity.

Languages are also different on a multitude of structural levels, which is referred to as structural diversity. For example, languages can differ in the way they order constituents, how they organize the sound system or how they code epistemic elements (Nettle 1998; Gavin et al. 2013). Nettle (1998) also speculates that structural diversity tends to correlate with phylogenetic diversity, since areas with a high number of language families also have a high amount of structural diversity. Nichols (1992: 250) has ended up with a somewhat similar conclusion in her study by claiming that "high genetic density and high structural diversity coincide in their geographical distribution". The structural diversity is the focal point of this thesis, since I will compare a set of languages by measuring their structural differences.

By now it is clear that the world is highly diverse linguistically. However, the global distribution of diversity is uneven across continents. The reasons behind this skewed distribution are obviously manifold, and despite the recognition of this issue, the knowledge of the mechanisms affecting language diversity is still very limited (Gavin et al. 2013). As seen in Figure 1, the diversity density is clearly substantially higher near the Equator, mostly between the two Tropical Circles, Cancer and Capricorn. There are two great belts of density in the area between the Circles: one within Africa, running diagonally from the West Coast to the East Coast, and the other around Southeast Asia, spanning from India to the Pacific (Nettle & Romaine 2000). According to Nettle & Romaine, these areas contain approximately 60 % of the world's languages, but constitute only 9 % of the world's land area and only 27 % of the world's population, emphasizing the unevenness of the distribution.

Inevitably, these two density hotspots are also areas with a high amount of structural diversity. The linguistic diversity in South America is also quite substantial, as is clear from Figure 1. The choice of South America as my research area was motivated by its high diversity, and the fact that it has not been studied from this perspective. Additionally, South America is very diverse in the geographical distribution of languages, since the languages are highly discontinuous in their distribution, resulting in a web of intertwined languages. This intriguing combination of linguistic and distribution diversity makes South America definitely worth studying.

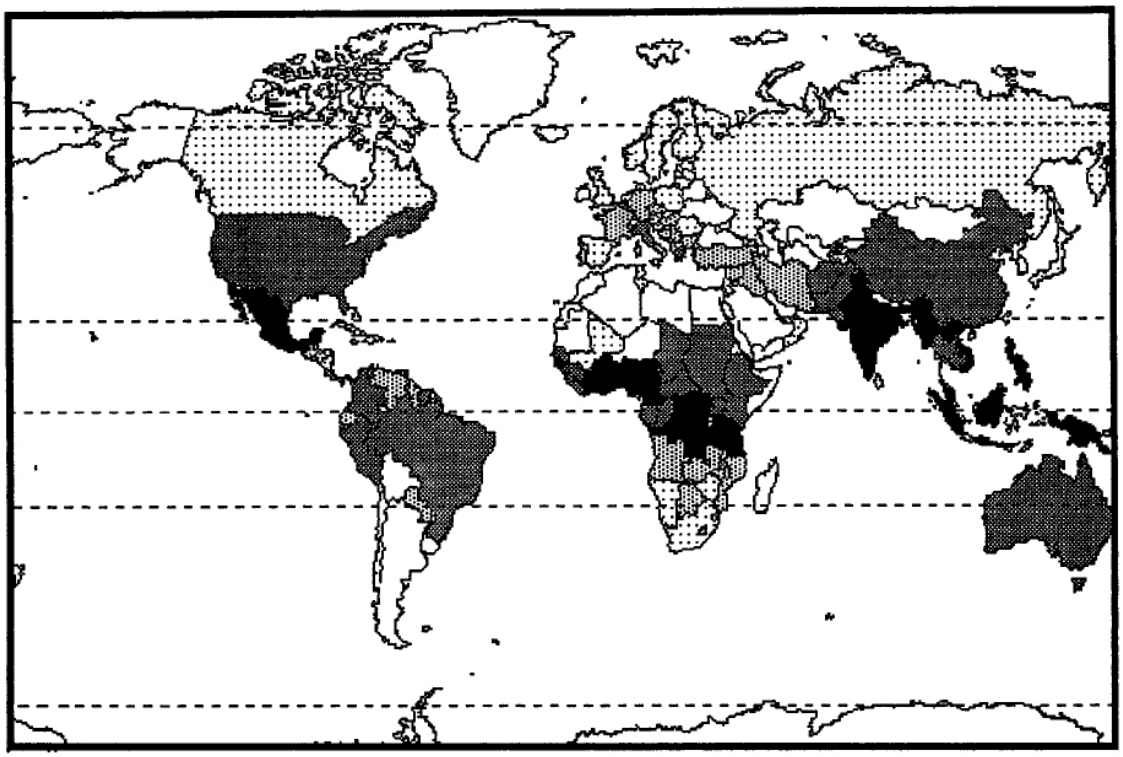

Figure 1: The density of language diversity (Nettle & Romaine 2000: 33).

2.1.2 Analysis of changes in diversity

Humankind has always been changing and developing, so it is clear that the evolution and development of the modern world has been through a lot of fluctuation. Linguistic diversity, like any other aspect of human existence, has also gone through a lot of changes and developments. One of the most famous models for the fluctuation of languages is R. M. W. Dixon's punctuated equilibrium model, which is based on a similar model used in biology. I will elaborate on this model and consider the different punctuations and the state of equilibria according to the framework of Dixon (1997) and Nettle (1999), which I will later on use in this thesis when explaining the changes in linguistic diversity in South America.

An important issue to address is the fact that neither Dixon or Nettle claim that societies and languages are completely stable for thousands of years and then suddenly change immensely. All languages are fluctuating all the time, even in an equilibrium situation (Nettle 1999). A rather exquisite metaphor for equilibrium by Nettle is comparing it to a gas: a gas is stable when its pressure and volume are stable, but the molecules in it are nevertheless moving constantly. This applies to equilibria as well. Languages and speech communities are stable when nothing drastic is happening, i.e. the situation is not punctuated, but changes are nevertheless happening constantly within languages.

2.1.2.1 Characterizing equilibrium and punctuation

In his book *The rise and fall of languages* (1997), Dixon presents a hypothesis about the development of languages and linguistic diversity. He hypothesizes that human history has mostly been stable, but this stability has gone through several different interruptions. Dixon describes these fluctuations by using his punctuated equilibrium model, which is based on two key concepts, punctuation and equilibrium. These two concepts refer to a break or interruption and to a state of balance, respectively. Different punctuations can disturb the equilibria, usually in a severe way. As Dixon states, these punctuations "give rise to expansion and split of peoples and of languages" (1997: 4), so whatever the punctuations are and whatever causes them, they result in dramatic changes within languages and between languages.

Dixon suggests that the human language has been in a state of equilibrium longer than the punctuations have lasted, the most recent equilibrium still being in effect. Nettle (1999), on the other hand, specifies an even more current punctuation caused by the Industrial Revolution (see § 2.1.2.2). In Dixon's model, the linguistic equilibrium can be described by the features societies speaking a given language are prone to have. In an area large enough to maintain several communities, each group is a political group with each of them having their own distinct language or dialect, and a name for that vernacular. When comparing these groups, no single community would be relatively larger in terms of population size than the others, and the communities

would also have somewhat similar lifestyles, beliefs and traditions. An equilibrium status is further enhanced by the lack of prestige, which means that no group and no language would have greater prestige than the others. What is emphasized here is the overall state of equilibrium. As mentioned above, changes are happening constantly within languages. However, the changes during an equilibrium would be rather minor compared to the changes caused by punctuations. During an equilibrium languages would diffuse with each others, becoming more similar, while punctuation would cause languages to diverge from one another.

But what is causing these punctuations in the first place? According to Dixon (1997), punctuations are usually affiliated with population growth. As mentioned above, languages would diverge due to the expansion of communities into new ones, which would eventually cause the languages to split. This process would continue at a steady rate, resulting in the development of new languages. When considering the state of linguistic equilibrium in general, there are both linguistic and non-linguistic reasons for a possible interruption of that balance. The linguistic factors causing punctuations could be twofold: first, some language might have prestige due to certain grammatical features, making it more desirable to communicate with, and second, one language might achieve a communicative advantage due to borrowing from another language. Either way, the result is the same. These so-called better-equipped languages would eventually split into new languages, making sure other languages were on their way to extinction. In a way this process could be seen as the survival of the fittest, the natural selection of languages. Obviously these linguistic triggers, as Dixon states (1997: 77), "are simply speculations".

Most punctuations, however, are due to non-linguistic factors. Dixon recognizes two main non-linguistic parameters, the causal and the geographical parameters. The causal parameter consists of natural causes, material innovations, aggressive tendencies and forms of communication, and these could all lead to a period of punctuation. Natural causes are related to the environment: some kind of natural disaster, such as a volcanic eruption or a flood, affects the living conditions in an area, forcing people to relocate. Some populations might even perish completely. Another natural cause of punctuation is the spread of infectious diseases, which results in a decrease in population, since people might lack the immunity to withstand certain diseases. But as people develop, so do their skills and abilities.

During human history, several major innovations have affected the equilibria, one of the biggest ones being the development of agriculture, which meant the spread of farming and a rapid population growth. The invention of tools and weapons inevitably made some groups have more advantage over others. The tools might spread to other groups, or it may lead to one group wiping out others with guns, and those without firepower would gradually vanish alongside their language. Another important innovation was the development of transportation, which facilitated the relocation of communities into new areas. By "aggressive tendencies" Dixon means the rather drastic measures of conquering new areas and oppressing peoples. A certain

group might have the required methods, e.g. weapons, tools or just a majority of the people, and through these it would gain power and prestige. This group would also have a prestige language. The remaining non-prestige groups would lose speakers because of the higher status of the prestige language. People would switch to that language and eventually, there would be no one left to speak the non-prestige language.

The most important innovation according to Dixon is the development of writing and other forms of communication. He divides the reasons for its importance into two, seeing writing as generally beneficial and selectively beneficial. The former refers to the general importance of writing, enabling people to communicate and develop literature, while the latter means how most writing is published in the prestige language of a nation, making the non-prestige language speakers abandon their languages. In a way writing is seen as a privilege of large powerful groups, which leads to smaller, usually indigenous groups to only gaining access to written materials in English or Spanish, for example. This same kind of imbalance has continued with the invention of radio and even television, which are mostly broadcasted in prestige languages, affecting the local languages.

The second non-linguistic parameter causing punctuations is the geographical parameter, which Dixon (1997) has divided into three separate possible sub-parameters: (1) the expansion into uninhabited territory, (2) the expansion into previously occupied territory, and (3) the confinement within a geographical area. As Dixon states, the expansion into uninhabited areas might happen due to several different reasons, for example when acquiring a more attractive place to live in. Whatever the reason, the result is the same, a punctuation. The new inhabited territory facilitates population growth, which results in newly formed groups of people and thus new languages. While the expansion into uninhabited areas probably leads to language splitting, the exact opposite situation occurs when a group of people expands into previously occupied territories. This expansion interrupts the equilibrium, because the invaders are more likely to have dominance over the existing population, making the invaders' language the prestige language. Eventually the original languages will either decline in use entirely or remain in use, but not as a prestige language. This type of expansion has been happening all over the world ever since the European migration started in the 15th century.

A certain punctuation might originate within a specific area. One example of this type of punctuation comes from South America, which is the research area in this thesis. Dixon states how the introduction of agriculture caused a punctuation within the Amazonian forests. The inhabitants of that area were hunter-gatherers speaking a variety of languages. When agriculture was developed, probably by one tribe as Dixon suggests, it eventually started to spread into neighboring tribes. This spread of agriculture caused a punctuation where populations expanded and languages split, which had an interesting consequence on the continent's linguistic diversity. On the one hand the area's hunters-gatherers were replaced or absorbed by the agricultural tribes,

and on the other hand the expanding speech communities resulted in their languages splitting into new ones over time. This describes the ebb and flow of the model quite well. The general features of South American linguistic diversity are described more thoroughly in § 3.2, while the explanations for the changes in the diversity are discussed in § 6.2.

2.1.2.2 Fluctuations in linguistic diversity

This section focuses on outlining the most important periods of punctuation which essentially affected linguistic diversity in general. In a way, our history can be analyzed by using the punctuated equilibrium model. I base this section mostly on the adaptation seen in Nettle's work (1999), which is, to be explicit, obviously somewhat speculative since some aspects of human history can only be roughly estimated. In his work Nettle provides short descriptions of a Palaeolithic equilibrium and of two massive punctuations, the Neolithic and the Industrial punctuations. The Palaeolithic era refers to the period before 10,000 BC, when the earliest stone tools were made, and the Neolithic period, occurring right after the Palaeolithic, indicates the development of agriculture (Bowens 2011).

By the Palaeolithic equilibrium Nettle (1999) refers to a relatively stable situation across the world, in which humans had been gradually spreading across the present-day continents. He also mentions how languages already existed during the latter period of this era. Humans were hunter-gatherers, and they were living in quite small societies, ranging from a few hundred to a few thousand people, consuming the resources available in a given area. After the resources were exhausted, they had to relocate, which was much easier with a smaller community, so the societies never grew too large. During the Palaeolithic era language diversity was probably increasing as was the population. Estimating the number of people and even estimating the number of languages is of course rather hypothetical, but Nettle uses the Australian amount of approximately 1,000–3,000 people per language to conclude that there might have been around 1,500–9,000 languages at the end of the Palaeolithic era, right before the first big punctuation.

As mentioned above, the Neolithic period started when humans shifted from hunting and gathering to agriculture. This resulted in the first big punctuation after the relatively stable Palaeolithic era, and this spread of farming affected a lot of things. Even though diseases also spread, the rise in birth rates outweighed the mortality rates. This led to a rapid population growth, which then resulted in migration, because the growing population needed more space. Some languages spread vastly, while some split into other languages, whereas others had the more upsetting fate of being replaced or even dying.

This punctuation had a massive impact on people and their languages, though Nettle suspects that the Neolithic period and its consequences for language diversity continued throughout the upcoming several millennia. Nettle calls this the Neolithic aftershock. The exponential growth of the population resulted in a number of big languages in Eurasia in terms of the number of

speakers. The nations which spoke those big languages had armies, officials and writing systems, which were all developments agrarian communities lacked, which in turn meant that these huge languages had a way of spreading. Due to the population growth, there was a lot of pressure in Eurasia to expand beyond its borders, which evidently led to several well-known voyages, one of the most famous being that of Christopher Columbus. What started then in the 15th century resulted in a 400-year-long expansion and migration of Europeans all across the world. This meant that the indigenous inhabitants were either killed, enslaved or banished from their territories. If someone was not killed or enslaved, they probably died of an infectious disease, eventually[2]. It goes without saying that linguistic diversity was dramatically affected by the mass depopulation of indigenous peoples and the expansion of European hegemonies. The main difference between the Neolithic period, the development of agriculture, and its aftermath is their actual effect on language: whereas the Neolithic punctuation meant the spread of agriculture and thus the gradual development of smaller farming communities which enabled language diversification, the European expansion did not cause divergence, but mainly the loss of language diversity.

Nettle recognizes an even more recent punctuation which he calls the Industrial punctuation. The Industrial Revolution, starting in the late 1700s, caused major changes in the way people lived their lives. Societies, mostly Western ones, became richer, healthier and more advanced, which led to a situation where languages were dying even though people were not moving. People were shifting to other languages within their habitats, since those languages had higher socioeconomic prestige. This pattern has caused a massive extinction of languages, making most of them endangered (Nettle & Romaine 2000). In recognizing these different punctuations, Nettle stresses the fact that these fluctuations of equilibrium and punctuation cannot just be seen as single processes in time. As mentioned above, languages are never in a completely stable condition, but react to every aspect of human existence. However, the recent punctuations have been drastic to language diversity. Nettle & Romaine estimate (2000) that at least half of the world's languages will become extinct during the next century. Nettle's hypothesis is a future equilibrium where only one or two languages exist, since the languages of the developed world are spreading rapidly.

According to Nettle (1999), the Industrial punctuation is still affecting languages. For example, the South American forests are being demolished, forcing indigenous tribes to relocate. Additionally, the speakers of indigenous languages are shifting to the hegemonic languages of Spanish and Portuguese due to socioeconomic reasons. This punctuation has definitely affected the linguistic diversity in South America, so I will use it as one of the ways of explaining in detail how the continent's diversity has fluctuated.

[2]For example, up to 95 percent of all Native Americans died from successive epidemics (Nettle & Romaine 2000: 117).

2.2 Theoretical approaches

In this thesis my aim is to explore the structural diversity of languages spoken in the tropical Andes and in the Peruvian Amazon by measuring their linguistic differences. I will also suggest explanations for these differences by paying attention to the geographical and social environments in which the languages are used. The following sections compile the language-ecological approaches used to explain the linguistic differences in the research area. I will first specify different approaches for studying linguistic differences, which is then followed by descriptions of language ecology and contact-induced language change.

2.2.1 Approaches for studying linguistic differences

The linguistic diversity of the world is visible not only by the sheer number of natural languages, but also within the languages themselves. All languages are complex entities, some more than others, but they all differ in vocabulary, grammar, written form, syntax and in other characteristics (Chiswick & Miller 2005). Scholars have wanted to group languages for centuries based on their differences and similarities, which has been the basis for historical linguists in order to classify language families. It is well known that historical linguists can reconstruct the prehistory of languages and determine their relatedness by using massive amounts of data, but it is all based on the notion of classifying languages as more or less similar by comparative examination.

The differences in languages can demonstrate how close or how far some languages are from one another linguistically. The result is the concept of linguistic distance, which refers to the "extent to which languages differ from each other" (Chiswick & Miller 2005: 1), whether between languages, language varieties or idiolects. It is quite impossible to measure linguistic distance the same way as measuring the distance between two countries, but the idea of this type of distance can be used, for example, to measure the ease of learning a new language (see Schepens et al. 2013). Because languages are very complex, linguistic distances could be measured according to the differences in phonology, syntax (i.e. in syntactic typology) and semantics (Nerbonne & Hinrichs 2006).

Borin (2013) notes how there has recently been a growing interest within the field of computational linguistics to actually measure these linguistic differences, focusing on measures that can be computerized and then automatically applied to large data sets. Essentially the differences between two language systems can be estimated, and then put on a numerical scale. This results in a linguistic distance measure, which should desirably be "metric in the mathematical sense, i.e., that the distances are non-negative and symmetric" (Borin 2013: 20). He states how linguistic distance measures can help linguists to group languages, and because there are many ways to group languages, there are also a lot of potential distance measures to be used to study differences between languages.

An important aspect of measuring differences is to decide what to actually investigate and in the end compare. Borin (2013: 9) classifies three different sets of concepts which can be compared when calculating linguistic distances:

(1) Sets of features
(2) Probability distributions of linguistic features
(3) Symbol sequences

The first concept refers to a set of either grammatical, lexical or phonological features, such as phonetic feature n-grams or part-of-speech tag sequences, while the second concept refers to the likelihood of certain features existing in the first place. The third concept basically means that words of the given languages are measured, usually by string similarity measures. Any of these three concepts could be used when calculating linguistic distances, but Borin concludes that the most used linguistic data has actually been basic vocabulary lists. He mentions how comparative-historical linguists tend to favor structural features when comparing languages, but adds that there does not seem to be a well-justified basis for the use of either structural features or vocabulary lists.

Chambers & Trudgill (2004) propose an order of importance for linguistic features when comparing languages. Even though their suggestion is focused on dialectological comparison of languages, especially the grading of isoglosses, I believe this ranking (Chambers & Trudgill 1998: 99) represents the structural significance of linguistic features quite well:

LEXICAL	1. lexical
	2. pronunciation
PHONOLOGICAL	3. phonetic
	4. phonemic
GRAMMATICAL	5. morphological
	6. syntactic

Here we can see how Chambers & Trudgill evaluate lexical items to be the least reliable and grammatical features to carry the most weight when comparing languages. Nerbonne (2003) has reached a similar conclusion when stating that lexical elements are less consistent than other linguistic features and hence more unstable, especially when comparing between languages.

Based on the representation of features I decided to follow this ranking when choosing what features to compare in order to measure the linguistic differences between a set of languages. Because Chambers & Trudgill suggest that grammatical features are the most reliable data on languages, I will use them as the base for the language data parameters, which are discussed more thoroughly in § 4.2.

2.2.2 Ecological view on language

As mentioned above, I intend to suggest possible explanations for linguistic differences in this thesis. When looking for explanations for the differences between languages it is important to remember that languages are not secluded entities existing independently from people, societies and the environment, but that there exists a connection between languages and the environment (Wendel 2005). I will base the explanations on a field of linguistics called language ecology, which is a holistic, multi-faceted and dynamic perspective that studies the interaction between language and its environment (Eliasson 2015; Haugen 2001). A Norwegian-American linguist Einar Haugen formulated this subfield of linguistics in his essay *The ecology of language*, because he felt that there is a need for research, where concepts of ecology are applied to language. For Haugen the environment meant first and foremost the society where languages were used on a daily basis, but as de Busser (2015) states, the scope of Haugen's formulations is not entirely clear. In general, it is important to study the relationship between languages and the ecological environment, as scholars such as Gavin & Stepp (2014), Gorenflo et al. (2012), Nettle (2009) and Axelsen & Manrubia (2014) have shown.

However narrow or wide the scope of Haugen's original formulation of the ecology of language is, the main reason for the use of an ecological approach in this thesis is to offer ways of explaining linguistic differences between languages by paying attention to the overall connection between language and the ecological environment. This approach is motivated by de Busser's view on how "linguistic structure is formed, changed and influenced by different aspects of the human environment" (2015: 1). In his view these different aspects can either be internal or external factors, which both influence languages. De Busser lists six external factors: cultural factors, social factors, geographical factors, natural factors, human biology and the meta-perception of language. De Busser differentiates between geographical and natural factors, because he sees that the two are considered to entail different aspects of the environment. He describes factors such as the physical proximity and the latitudinal gradient of languages as geographical factors, whereas natural factors are elements of the natural environment, such as rivers, mountains, flora and fauna, adding that both of these factors can cause language diversification.

To simplify, the geographical parameter used in this thesis to explain linguistic differences will include both the geographical and natural factors from de Busser's classification. This way the parameter will cover the widest possible range of explanations. In addition to the geographical factors I will also base my explanations on socioeconomic factors, which correspond to de Busser's social factors. In this section I will further elaborate on the geographical and socioeconomic aspects of the human environment based on de Busser's work. However, I will not use cultural factors, human biology or the meta-perception of language as explanations of linguistic differences, because they are not the central factors considering the scope of this thesis.

The social factors relate to the demographic and socioeconomic structures found within societies, such as power relationships, ethnicities, community sizes, interactions between other societies and social networks (de Busser 2015). In general different social factors can affect language use even without people noticing it, since speakers can change their language according to whom they are talking to, what the topic and the function of the discussion is, and what the overall social context of the discussion is (Holmes 2013). For example, people start shifting from one language to another in order to gain a better socioeconomic status. This shift usually derives from the desire to acquire better education and employment. A demographic transformation has caused the division of rural and urban areas, because more people are living in large cities than in rural areas (Veblen et al. 2015). Language shift understandably affects the small rural languages, because they are not seen as useful as larger urban, usually hegemonic, languages, so the smaller languages lose speakers due to these socioeconomic factors (de Busser 2015).

The geographical factors involve elements in the natural environment, which can influence languages and even their diversity and complexity (de Busser 2015). Examples of geographical factors in play are shown in studies which demonstrate how geographical proximity can correlate with both species and language diversity (Mace & Pagel 1995; Kerr & Packer 1997). A homogeneous area can lead to contact between languages, while a heterogeneous area can promote language isolation, which is strengthened by the vicinity of oceans and mountains. Because the social interaction with neighboring groups becomes more difficult, it can affect the spreading of languages and thus language diversity. There are nevertheless mixed results on the strength of this correlation, as Gavin et al. (2013) mention.

An interesting example is the relationship between Rapoport's rule and language richness by Gavin & Stepp (2014). Rapoport's rule is an ecological pattern which postulates that "the geographical extent of species ranges increases at higher latitudes" (Gavin & Stepp 2014: 1). This means that at higher latitudes the geographical area where a given species can be found increases in size. Gavin & Stepp studied if analyzing Rapoport's rule might help understand the causes behind the skewed distribution of language diversity. They conducted a global analysis in order to analyze the magnitude of Rapoport's rule, and they found out that the distribution of language range, i.e. the area where languages can be found, is skewed. 87 % of languages Gavin & Stepp studied had a range area less than 10,000 km^2, while only a few languages had a range area over 1,000,000 km^2. They point out how language richness correlates with latitude, meaning that there are more languages in the tropics than at higher latitudes. At the same time language richness is negatively correlated with language range, so when one increases, the other decreases. Gavin & Stepp hence found evidence that Rapoport's rule influences latitudinal diversity on a global scale. There are obviously many processes affecting language diversification, but the strong latitudinal aspect in language diversity and language range sizes imply that there are also some environmental components in effect.

Geographical factors can also be smaller scale items, such as the influence of river systems and mountains, or large scale relationships, such as the relationship between biological and linguistic diversity (de Busser 2015). Several scholars have noted how linguistic and biological diversity occur relatively often in the same regions. Patterns of co-occurrence can be identified between these two diversities in areas such as West Africa, Melanesia, Mesoamerica and especially New Guinea (Gorenflo et al. 2012). Languages in these areas account for approximately 70 % of all languages, and most of them are usually endemic, i.e. native, to certain regions (Gorenflo et al. 2012). In Figure 2 one can see a representation of the number of languages within every biodiversity hotspot or wilderness area. According to Myers et al. and Mittermeier (2000 and 2003, both cited in Gorenflo et al. 2012: 8032), biodiversity hotspots are areas "characterized by exceptionally high occurrences of endemic species and by loss of at least 70 % of natural habitat", whereas biodiversity wilderness areas are regions of at least 10,000 km^2 "having lost 30 % or less of their natural habitat". In the 40 hotspots or wilderness areas identified in Figure 2, there are 4,824 languages spoken, of which 3,474 are endemic to the regions where they occur. Therefore it is quite clear that these biodiversity hotspots or wilderness areas are also linguistically diverse.

Correlations between linguistic diversity and biological diversity are also mentioned in Moore et al. (2002), who observed how biodiversity can speed up the diversification of languages through

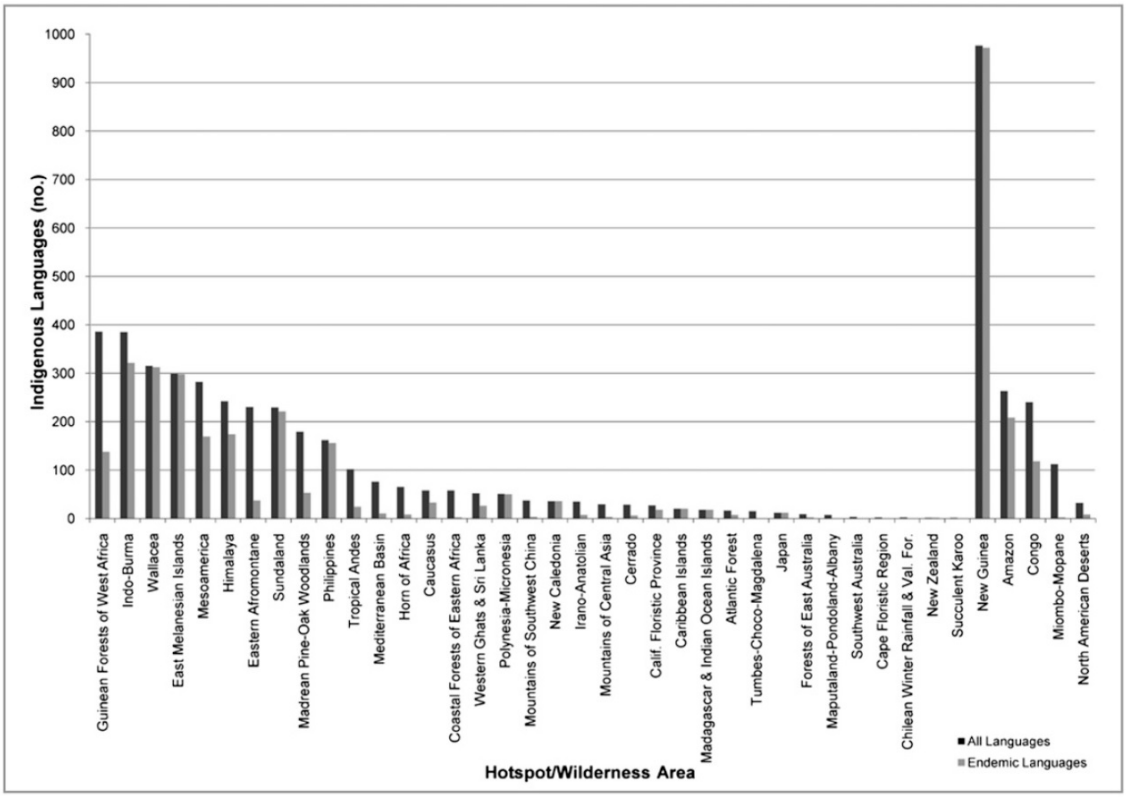

Figure 2: Distribution of languages within biodiversity hotspots (Gorenflo et al. 2012).

resource partitioning, which indicates the way species avoid competition for resources (Dunn 2003). Evading competition can mean anything from sharing to dividing of methods, property, etc. Moore et al. also note how fragile both linguistic diversity and biodiversity can be; they can both be destroyed by social processes and new technologies, and they both react to environmental factors in similar ways. However, it should be mentioned that even if language diversity and biological diversity correlate with each other, that does not necessarily imply a causal relationship between them (Gavin et al. 2013).

Overall the external geographical and socioeconomic factors discussed in this section have demonstrated the complex relationship between the environment, languages and their speakers. This ecological view is an acceptable approach in explaining linguistic differences, because as Lupyan & Dale (2015: 294) claim, "languages adapt to their environments". If languages are spoken in similar ecological environments, they tend to become more similar, and vice versa; different social structures and physical environments will facilitate linguistic diversification. In general the field of language ecology emphasizes how the interaction between language and its environment is worth studying and can result in new discoveries.

2.2.3 Contact-induced language change

I am exploring linguistic differences in this thesis, but it is important to remember the other side of the coin, which is linguistic similarity. The two phenomena seem to be contradictory to one another, but they are in fact related, because they are both connected to language contact. Fill (2007: 179) states how "language contact is only possible because of language diversity". Because there are so many languages, mostly in dense diversity hotspots, languages tend to interact with other languages in the same area. This leads to a process of contact-induced language change, when genealogically unrelated languages, which previously had significant linguistic differences, start to resemble one another (Winford 2005). On the other hand, genealogically related languages may become different through language contact. In an extreme case, language contact may lead to language shift, so therefore language contact is a noteworthy approach when exploring structural similarities and differences, because it affects the overall linguistic diversity. On its own language contact is an extremely large field of study, so in this thesis I will only briefly mention some of its postulates, which are relevant for the present study.

According to Lucas (2015), there are a lot of controversies regarding the definition of language change and the role of language contact. Lucas specifies how there are basically two opposing views on this matter: the first one, favoring Chomsky, defines language change as happening on an individual level, while the second one, favoring sociohistorical approaches, defines it as happening to all speakers of a group. Both of these views differentiate contact-induced change from language internal change, because in the former the change is due to bi- and/or multilingualism

in the speech community, and thus due to language contact (Lucas 2015). Nevertheless, these two opposing views do not agree on what contact-induced change actually encompasses. As Lucas states, for the individualists contact-induced change indicates a change within a speaker's idiolect, which differs from the grammars of older speakers, whereas the sociohistorical perspective sees the change occurring when all the speakers of a group have changed their linguistic habits. In general language contact is regarded to concern young people more likely than old people (Heine & Kuteva 2005), which in a way supports the latter sociohistorical perspective of group level change. The main purpose of this approach is to offer plausible ways of explaining linguistic differences between a set of languages, so I am interested in the contact-induced changes on a group level. However manifold the definitions and the reasons for language change are, the most important fact to keep in mind is the entire phenomenon of language contact, which happens when people change their linguistic tendencies after being exposed to other languages (Heine & Kuteva 2005).

In their seminal work Thomason & Kaufman (1988) present two types of contact-induced language change, borrowing and interference through shift. They have described borrowing as "the incorporation of foreign features into a group's native language by speakers of that language" (1988: 37), whereas interference through shift happens when "a group of speakers shifting to a target language fails to learn the target language perfectly" (1988: 39). However, as Winford (2005) states, Thomason & Kaufman do not clearly explain what these terms actually mean and include, since they are just regarded as processes of language change without further explanations. Overall, language contact can be viewed as a borrowing process of any kind of linguistic feature (Winford 2005), while language shift occurs when speakers of a language shift to another language. Usually the shift is voluntarily and due to social factors, such as people's desire for education and better employment (Holmes 2013). Language shift has a reducing effect on linguistic diversity, because people are speaking fewer languages due to the shift.

The exposure to other languages and its consequences is the crucial linguistic factor when explaining similarities and differences between languages. Even though some linguists still claim that the only proper account of contact-induced language change is the existence of loanwords, I justify the use of language contact as an approach by citing Thomason & Kaufman, who state how "any linguistic feature can be transferred from any language to any other language" (1988: 14). Therefore language contact explains not only lexical, but also structural similarities and differences between languages. These structural features are the focus of this study, which will be examined later on in Chapter 4. Overall, languages can diversify and become similar in many different ways, since in addition to linguistic parameters also geographic and socioeconomic parameters affect languages (Aikhenvald 2006b). All the aforementioned factors are used to explain the structural differences between the languages studied in this thesis.

3 South America: diversity, geography and demography

South America is a continent showcasing remarkable linguistic diversity. But as mentioned in the previous chapter, languages do not exist in seclusion, but as an essential part of societies. Therefore it is appropriate to further examine the entire continent in order to fully understand the environment where these languages exist. In this chapter I will first focus on describing the possible ways South America might have been inhabited, which is followed by a summary of the continent's overall linguistic diversity. Subsequently I will portray South America's geographical and demographical features. In addition, because the languages compared in this thesis are spoken in the tropical Andes and the Peruvian Amazon, encompassing the countries of Colombia, Ecuador and Peru, I will further elaborate on these countries in every section. None of these descriptions are meant to be exhaustive, because my focus is to provide an overall representation of these features, so that the results and the explanations in this thesis can be easily situated into a proper context.

3.1 How the Americas were inhabited

In order to fully understand the processes behind the formation of linguistic diversity in South America, it is important to hypothesize about the peopling of the Americas. Human history is full of interesting details and happenstances, and one of them is the complex phenomenon of how early humans migrated to North and South America. There are as many theories for the peopling of the Americas as there are studies on the subject. In this section I will elaborate on the theories and thoughts of Nichols (1990), Goddard & Campbell (1994), Blench (2008), Mulligan & Szathmáry (2017) and Peterson (2011), of which the first three are based more on linguistic data than the two remaining ones, which focus on other kinds of data. Throughout this section I will use the term New World when referring to the American continents, as it is a well-established and widely used concept.

The 1990 paper from Nichols is a thorough description of the first settlement of the New World

by using linguistic data in order to estimate the age of linguistic populations. Nichols aims to strongly criticize Greenberg's 1987 study, where he suggested a time frame for the first migration into the New World being around 12,000 to 20,000 years ago. Nichols tries to verify, at least on some level, if these estimates are even somewhat accurate or completely wrong by estimating a more accurate age for the language families. These estimates are based on the linguistic diversity and on the linguistic differentiation rates found in the source area, which was northeastern Siberia. She agrees with other scholars when stating that there were probably multiple entries into the New World, and that these entries were made by people speaking the same language or closely related languages. She concludes that based on the evidence from her studies on linguistic prehistory, it seems more likely that the New World has been inhabited for tens of millennia, longer than generally thought. Even though her paper was interesting, Nichols seems to focus mainly on criticizing Greenberg and proving him wrong, so she does not really comment on the debate on where the first populations of the New World originated from and how did they come there. For her the debate has been mostly about the age question, since previous suggestions have been inconsistent with linguistic facts in her opinion.

Goddard & Campbell continue on the same path as Nichols when criticizing Greenberg's opinions, stating that it has way too far-reaching a scope. They feel that considering the complexity of human linguistic history, it is not clear how it corresponds to non-linguistic history, but it might end up shedding light on the matter of peopling of the Americas. In their view, several possible scenarios for the migration of the people and their languages into the New World might be plausible. The suggested scenarios are as follows:

(1) A single migration occurred at some point in time, and later this genetic unit diversified, forming several language families.
(2) Populations in Northeast Asia and their languages underwent a differentiation, and these distinct units traveled to the Americas over time.
(3) There were several different migrations all of which happened during separate time periods and consisting of separate languages.
(4) A single migration occurred, but there was more than one language present.
(5) A possibility that one or more of the language groups which migrated ended up going extinct.

After presenting these different scenarios, Goddard & Campbell conclude that the relationship between linguistic and non-linguistic history is a complex one, making these suggestions of scenarios quite risky. According to the authors, the most important aspect to keep in mind when constructing theories for migration is the fact that there is more linguistic diversity in the Americas than in Eurasia, because the Americas were populated relatively recently.

The third linguistic paper is from Blench, who gives a very specific account on the linguistic

diversity of the Americas, focusing on explaining language groupings and language expansions. In the end Blench states that physical anthropology should be used more often alongside genetic modeling of populations to fully understand the peopling of the New World. In his hypothesis hunter-gatherers migrated first to the New World around 25,000–30,000 BP[1] from Siberia, and those smaller groups might have eventually spread out down the West Coast. These migrations would continue across Beringia, through which diverse language groups would find their way to the area. Ultimately local hunter-gatherer groups would keep on growing and expanding, and sooner or later also diversifying. Blench seems to base his hypothesis rather optimistically on the anthropological findings, but he nevertheless suggests that the New World has been inhabited for tens of millennia.

The remaining two theories are not based merely on linguistic data, but on other scientific aspects that studies have shown to justify. The first of these is the study made by Mulligan & Szatmáry, where they modeled the peopling of the Americas by using molecular genetic data. The hypothetical model based on this type of data depicts that a population originating somewhere in Asia diversified around 40,000 years ago, migrated to Beringia, where it existed in isolation, hence diverging even more. Then approximately 16,000 years ago there was a single rapid expansion from this isolation into the Americas. Mulligan & Szatmáry favor this model because the genetic evidence indicates that there are similar groups of DNA found throughout North and South America, but not in Asia. This would mean that these specific variants have emerged before any groups of people migrated to the New World, implying that the genetic variants have formed after the divergence from the original population in Asia. Mulligan & Szatmáry also managed to suggest a duration for the Beringian isolation by using the same molecular data. They estimated that the time needed for genes to diversify would range from 7,500 years up to 15,000 years. This new and interesting method seems believable since studying DNA is not based on facts unbeknown to us, but on scientific proof.

The second nonlinguistic theory comes from Bennett Peterson, who focuses on discussions of migrations to the New World by sea. She states explicitly how she accepts the theory of early humans crossing the Bering Strait land bridge in order to migrate to the New World, but wants to distinguish other possibilities as well. The Pacific Rim Model is a model for the settlement of the Americas by sea, suggesting that ancient people were skilled seamen able to reach the Americas by boat. According to Bennett Peterson and other scholars, these ancient mariners were able to successfully complete their journeys by eating seaweed and relying on other resources as well. Due to glacial ice the most probable route into the New World would have been down the Pacific Northwest Coast, and some groups might have kept on moving alongside the coast as south as possible. Bennett Peterson also states that because of the ocean currents and their directions, groups of people might have migrated into South America from the south upwards already in

[1] Before Present

33,000 BC. It is clear in her description that she also believes that this migration into the New World is more ancient that previously thought.

In summarizing these five accounts on the peopling of the Americas it is obvious that there are many possible theories on this matter. Of the five authors mentioned here, all three, Peterson, Nichols and Blench, clearly believe that the migration into the New World started tens of thousands of years earlier than generally thought. It is a good thing that linguistic prehistory is used to support historical accounts, but it might be dangerous to view matters based only on linguistic history, since one might want to find some connections so desperately that the actual study and its results might not be objective enough. I chose these papers because they represent different point of views in their own respective ways, but I will not claim any of them to be the absolute truth and the only possible explanation for the peopling of the Americas.

Figure 3 illustrates the migration depicted in this section, including both possible land and sea routes to the New World. Even though the discussion on the inhabitation might sound like it is focusing mostly on North America, there is an understanding that this migration eventually also reached South America. Bennett Peterson briefly mentioned how some groups of people might

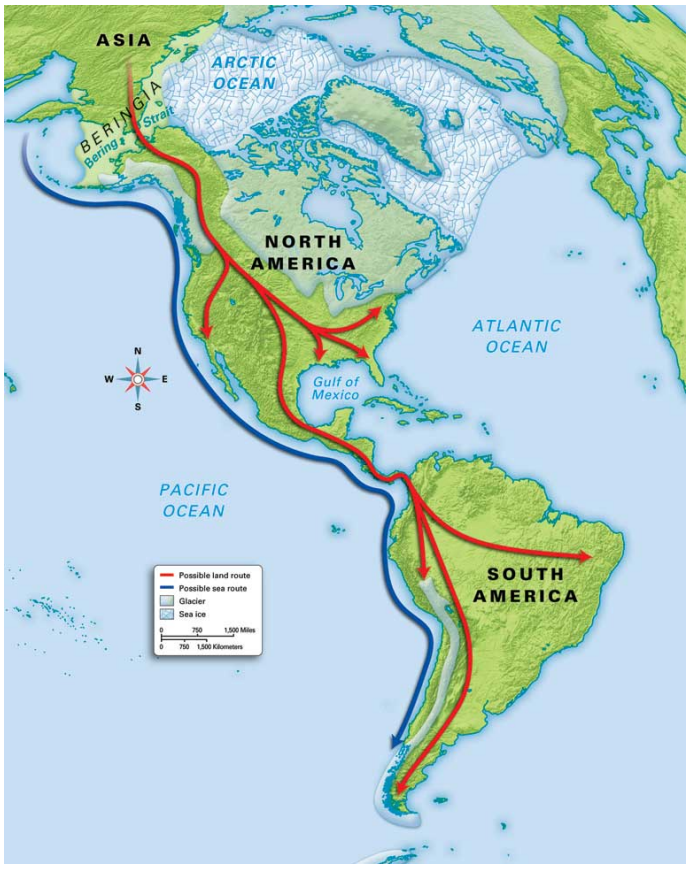

Figure 3: Possible land and sea routes to the New World.[2]

[2] http://mrgrayhistory.wikispaces.com/

have traveled to South America by sea along the South Pacific current, inhabiting the continent starting from the south. There are of course other views as well. Lynch (1999) mentions that the vast majority of anthropologists believe that the early humans who migrated to the Americas came through the Bering Strait land bridge, originating from Northeast Asia. From thereon, as these groups found their way south, they inhabited South America. According to the studies Lynch cites, it took the hunter-gatherers approximately 1,000 years to proceed from North to South America, ending up all the way in the southernmost parts of present-day Chile. It is rather impossible to describe how these people inhabited the areas in South America which are nowadays known as Colombia, Ecuador and Peru, but as people migrated into the entire American continents, they evidently settled in areas which later on became these nation states. Because people were mostly nomads, they moved from one place to another, so it probably took thousands of years before they settled down in certain locations, eventually becoming the forefathers for Colombian, Ecuadorian and Peruvian population.

3.2 Linguistic diversity in South America

As it was seen in Figure 1, South America is one of the linguistic diversity hotspots of the world. According to *The Ethnologue*, there are 455 living languages in South America, of which 136 languages are threatened and 137 are endangered. The astonishing amount of linguistic diversity in South America can be put into perspective when compared to the number of living languages in Europe, which tally up as 287 languages according to *The Ethnologue*. Even though there is a lot of information available on South American languages, some languages are described more accurately than others, while some are yet to be described at all, making them entirely unknown to us. Hence it is difficult to record the total number of languages. Several South American languages are also known by more than one name, which makes the classification of languages even more difficult. Campbell (2012) states various examples of languages with multiple names, one of which is the Nivaclé language, which has or has had nine other names. He also mentions (2012: 62) that the matter is further complicated "by the fact that [...] languages have come to be known by the name of the river [...] or by other salient geographical features".

Be that as it may, and despite the classification issues, there are approximately 108 recognizable language families in South America, 55 of which are language isolates (Campbell 2012). Dixon & Aikhenvald (1999) also describe how these language families each have their own characteristic cultural profiles, consisting of features such as the method of obtaining food, the material culture and the type of territory where the language exists. For example, the languages from the Arawak, Carib and Tupí families are found in the rainforests, and the population uses agriculture and they make canoes, hammocks and pottery. Some of the language families are not found just in small areas, but they have spread across a wide territory (Campbell 2012). As Dixon & Aikhenvald (1999)

state, this discontinuous distribution of South American languages results in a web of intertwined languages. One of these widely spread language families is the Arawak family, which as a language family contains the highest number of languages in South America and spans across eight countries. As with other linguistically diverse areas, there are some similarities and some differences between South American languages. Even though it is difficult to find out if the similarities are due to language contact or common ancestry, there are still a number of features occurring in several South American languages. Muysken (2012: 237) has identified the following features:

- Complex verbal morphology
- Agglutinative morphology
- Head marking
- Evidentials
- Both nominal and verbal classifiers
- Possession often marked on the possessed noun
- Clause subordination through nominalization

There are of course other features represented in the 108 language families as well, but together with the overall linguistic diversity, this area has understandably intrigued a lot of scholars over the years.

Dixon's punctuated equilibrium model described in § 2.1.2.1 can be used to describe the development of the continent's linguistic diversity. In general there are three distinct punctuations, which affected the languages and the overall diversity. Before the first punctuation, as Dixon & Aikhenvald (1999) specify, a period of equilibrium prevailed after the migration into the New World and the population of South America. The equilibrium probably occurred within each geographical area, whether the areas were rainforests or grasslands. The first punctuation was the development and adoption of agriculture, which also took place in South America. This resulted in the expansion and splitting of groups of people, which happened to the languages as well.

The second massive punctuation happened when the Europeans invaded South America from the end of the 15th century onwards. Indigenous people were either killed or enslaved, or died because of the spreading diseases. The invasion commenced the expansion of prestige languages, such as Spanish and Portuguese. This massive punctuation caused the inevitable extinction of South American indigenous languages – Dixon & Aikhenvald estimate that more than half of the languages originally spoken in South America have died. The aforementioned numbers from *The Ethnologue* tell the same sad story.

According to Nettle & Romaine (2000), the third punctuation was the Industrial Revolution, which led to diversifying economies but also inequalities in the acquisition of new technologies. This industrialization has also resulted in the loss of indigenous languages, because the Europeans wanted to exploit the natural resources of the continent, resulting in the demolition of the indigenous groups' habitats (Blouet & Blouet 2015). Nowadays even the remote tribes living in

isolation are not entirely safe from the European-style civilization, and slowly those languages are used less and less. Before all the indigenous languages disappear, it would be vital to document and describe these languages. This way the extraordinary linguistic diversity found in South America could be preserved.

Table 1: The numbers of languages in Colombia, Ecuador and Peru.

Country	**Number of living languages**	**Number of indigenous languages**
Colombia	84	79
Ecuador	24	21
Peru	93	91

Table 1 sums up the linguistic diversity in all the three countries of Colombia, Ecuador and Peru. It is clear that there are dozens of languages spoken in the three countries, of which a large percentage are indigenous languages. These indigenous languages consist of several language families and dozens of language isolates, which are, as it is common for the continent's discontinuous distribution of languages, spoken in various different areas. Despite the number of language families and isolates, there are a few language families which have a rather dominant position in these countries. In the Andean region, there are two languages with this type of dominance: Quechuan and Aymaran, which both have several million speakers (Adelaar 2012). The dominance is visible for example in Ecuador, where a local variety of Quechua, called Quichua, has replaced several languages and become the dominant language of the society and its people. On the other hand, there are also languages from other language families spoken in the Andean areas, the speakers of which have tried to survive the European conquest by residing in the mountainous area, forming linguistic islands (Adelaar 2004). Overall the linguistic diversity in these three countries is definitely interesting and well worth further study.

3.3 The physical geography of South America

South America as a continent is rather manifold. Its linguistic diversity is really substantial, but its physical geography is also diverse. South America consists of a wide range of diverse environments ranging from mountainous areas with altitudinal climates to tropical rainforests and temperate lowlands to deserts. The best-known feature of South America is its long mountain system, the Andes, which run all the way alongside the western coast of the continent. The Andes are actually the longest mountain system on Earth (Veblen et al. 2015). Blouet & Blouet (2015) state how the area west of the Andes consists of coastal plains, while the east side is mainly

Figure 4: The physical geography of South America (Veblen et al. 2015: 33).

plateaus and plains, but also highlands. They also define and separate four distinct areas in South America: the Andes, Amazon & Orinoco, Guiana Highlands and Pampa & Patagonia. All of these areas are categorized according to their physical geography, their environment and their climate. Figure 4 represents all these diverse topographic regions which constitute this massive continent.

The Andes are one of the most recognized features of South America due to its size and range on the continent. Most of the Andes are located in the tropical zone, but obviously the climate varies, since the high altitudes can modify both the climate and the vegetation. This also means that even in areas located on the Equator, there are snow and glaciers. Due to its position on the edge of a tectonic plate, the South American Plate, the Andes and its adjacent areas are prone to earthquakes and volcanic activity (Veblen et al. 2015).

On the other hand, the Andes are very rich in natural resources, such as oil, gas, gold, copper and tin (Blouet & Blouet 2015). These fertile lands are one of the reasons some mountainous areas have been inhabited for thousands of years. The coastal area between the ocean and the Andes varies in its terrain depending on the latitudinal location of the area. North of the Gulf of Guayaquil (located in south Ecuador, near the border of Peru) the coastal areas are mainly forests,

while south of the Gulf the areas are much drier and mainly deserts. For example, the entire west coast of Peru, so the area between the ocean and the Andes, is a desert.

The Amazon and the Oricono are massive rivers, which flow in the lowlands east of the Andes (Veblen et al. 2015). The lowlands usually have extremely high rainfall due to their tropical climate, but they also have a dry season. The Guiana and Brazilian highlands represent the oldest bedrock in South America, so they are remains of the eroded old mountain systems, which are a source of minerals to the inhabitants of the areas (Blouet & Blouet 2015). Additionally, similar to the lowlands, also the highlands have a tropical climate with a dry season. The Pampas is located south of the Amazonian lowlands and its fertile grasslands stretch all the way from the Atlantic to the Andes. The Patagonian plateaus are also located between the south Atlantic Ocean and the Andes. The vicinity of the Andes results in a cool temperate land with low precipitation since the Andes block winds coming from the Pacific, making the plateaus dry grasslands.

In addition to this division of South America into larger areas, Blouet & Blouet (2015) mention how the Andean countries, such as Colombia, Ecuador and Peru, can be divided into three distinct zones due to the effect of the Andes. These distinct areas are the Costa, the Sierra and the Oriente. The Sierra refers to the highlands of the Andes, and its surrounding areas, while the narrow coastal area between the Pacific and the Andes are called the Costa, and the wet interior lowlands east of the Andes are called the Oriente. The Costa alters between wet forests and dry deserts depending on the latitude. In Colombia and Ecuador, the Costa is mainly humid forests, but in Peru the entire Costa is a desert. The Oriente and its physical characteristics also vary depending on the latitude. Parts of Colombia's northern Oriente experience dry seasons, where as the areas near the Equator are tropical rainforests. From this description it is clear that in addition to South America's linguistic diversity, it is also diverse in its physical geography.

3.4 Demographics and socioeconomics in South America

According to the Population Reference Bureau (2017), the population of South America is 423 million (mid-2017). Of the thirteen nation states in the continent, Brazil has the biggest population with 207.9 million inhabitants, while Suriname is the smallest country by population of 0.6 million. In general South America is an interesting example of population distribution, because its population is not distributed evenly, but spread across the edges of the continent (Blouet & Blouet 2015). As (Blouet & Blouet 2015) state, the entire continent can be described as a "hollow" continent, since the interior regions have less than 10 inhabitant per square kilometer.

One contributing factor to this skewed population distribution is the physical geography of the continent. The interior regions are characterized by the Amazon basin, which, as stated in the previous section, is a massive river system. The presence of this kind of natural element and its rough conditions have affected the population density. On the other hand, the South American

landscape is also dominated by another massive natural element, the Andes. Despite its presence, populations have managed to inhabit the Andean territories for centuries. This is mostly due to the rich mineral resources the Andes provide (Blouet & Blouet 2015). For example, according to The World Factbook (2017), almost half of the population of Ecuador reside in the Andean basin.

In general, a major demographic transformation has occurred in all South American countries (Blouet & Blouet 2015). During the last century, people have migrated from rural areas to urban centers, which has caused a very rapid urban development. This transformation is visible in the population geography, since "South America has the highest concentration of population in large cities [...] of any of the world's continents" (Veblen et al. 2015: 324). This highlights the "hollowness" of the continent even further. The cause of this transformation according to Blouet & Blouet (2015) lies in the socioeconomic expansion of the countries. The economic growth enables cities to develop and prosper, which creates new job opportunities, causing the people from the provincial territories to migrate to the urban areas. This massive migration has emphasized the division between the rich and the poor, since there are drastic inequalities in wealth, income and education.

Economically South America has been characterized as focusing on the primary sector, which means exploiting natural resources, such as oil, gold, copper and coal, but also the cultivation of rich soils (Blouet & Blouet 2015). The continent is also well-known for its forest resources and fishing industries. According to Blouet & Blouet, there has nevertheless been a transformation in the primary section, since the mechanization of mining and agriculture has led to a decrease in the demand in labor force. Additionally, South American agriculture has become commercialized, meaning that more and more crops are being exported. During the second half of the 20th century the manufacturing sector gained momentum and steered several South American countries into excessive manufacturing production. For example, in 2012, Brazil produced 34,7 million metric tons of crude steel (World Steel Institute, as cited in Blouet & Blouet 2015). Another important element in South American economies is the service sector, and especially the tourist industry, which is a vital source of income for many societies and cities. The demographic transformation has also affected the economies, because the urban development has increased the production of manufacturing and service sectors, which in turn creates more jobs, which in turn accentuates the rural to urban migration.

Table 2 sums up some of the key demographic parameters of South American countries. Additionally, the three countries studied more thoroughly in this thesis are highlighted. This data pinpoints the two main features mentioned in this section, the skewness of the population density and the urban development. The third column depicts the population per square kilometer, which demonstrates the "hollowness" of the continent. In Colombia, Ecuador and Venezuela the densities are quite immense, since there are around 30–50 people living per square kilometer. All of these countries are located in the exterior regions of South America, which showcases just how

Table 2: Demographic data of South America.[3]

Country	Population mid-2017 (millions)	Population per km^2	Births per 1,000 people	Deaths per 1,000 people	Percent urban
Argentina	44.3	15.0	17	6	83
Bolivia	11.1	9.2	24	8	69
Brazil	207.9	23.8	13	7	86
Chile	18.4	22.3	14	6	83
Colombia	**49.3**	**41.0**	**18**	**6**	**76**
Ecuador	**16.8**	**52.9**	**20**	**5**	**64**
Guyana	0.8	3.5	21	8	29
Paraguay	6.8	15.9	21	6	60
Peru	**31.8**	**22.8**	**20**	**6**	**79**
Suriname	0.6	3.0	18	7	66
Uruguay	3.5	18.9	14	9	95
Venezuela	31.4	30.3	19	5	88

skewed the population density is. Additionally, the last column represents the percentage of people living in urban areas. From these statistics it is also clear how the demographic transformation has affected the South American demographies. From the twelve countries in Table 2 all but one (Guyana) have at least 50 % of their population living in urban areas. There are even five countries with a percentage of over 80, Uruguay being the most staggering example of urbanization, with 95 % of the people live in urban cities. These figures underline the drastic transformation that has occurred in South America. The areas I am interested in this thesis, Colombia, Ecuador and Peru, are all highly urbanized, even though the population of Ecuador is not as high as in Colombia and Peru. However, of all the three countries, Ecuador has the highest population density, which is 52.9 people per square kilometer.

As mentioned in § 3.2, there are 455 living languages spoken in South America. The majority of these languages are indigenous languages, which are spoken throughout the continent.

[3]The data was extracted from the Population Reference Bureau's World Population Data Sheet (2017), with the exception of the population density data, which is from `https://www.worldatlas.com/articles/south-american-countries-by-population-density.html`. Even though I mention thirteen nation states in the beginning of this section, the World Population Data Sheet does not have statistics on French Guyana, hence Table 2 has only twelve South American countries represented.

The indigenous languages are spoken by the indigenous population of South America, the South American Indians, who are an important part of the continent's demographics. The Indian population is present in most of the South American countries, but the distribution of the indigenous population increases when moving from north to south (Blouet & Blouet 2015). According to the Economic Commission for Latin America and the Caribbean, there are dozens and even hundreds of indigenous groups in South America (ECLAC 2014). For example, in Brazil there are approximately 300 different indigenous groups. In Colombia, Peru and Bolivia there are around 100, 80 and 40 different groups, respectively.

The specific demographics of the South American indigenous population are represented in Table 3, where the three countries of interest are again highlighted. There are three countries with a relatively high percentage of indigenous people out of the total population: Bolivia, Chile and Peru. For example in Bolivia, 62.2 % of the population belong to an indigenous group, which is quite astonishing. However, the percentages do not reveal everything about the total number of people. The percentage of indigenous population in Colombia is only 3.4 % but the total number of population tallies up to 1.6 million, which is almost as much as in Chile, the percentage of which is 11 %. So clearly the number of indigenous people in South America varies quite a lot depending on the country. Blouet & Blouet (2015) mention how indigenous statistics can vary due to different manners of gathering census data. In Colombia, for example, one can self-identify as

Table 3: Indigenous population in South America.[4]

Country	**Percentage of indigenous people**	**Total number of indigenous population**
Argentina	2.4	955,000
Bolivia	62.2	6.2 million
Brazil	0.5	900,000
Chile	11	1.8 million
Colombia	**3.4**	**1.6 million**
Ecuador	**7**	**1 million**
Paraguay	1.8	113,000
Peru	**24**	**7 million**
Uruguay	2.4	77,000
Venezuela	2.7	727,000

[4]Data extracted from ECLAC (2014).

indigenous, which resulted in people choosing not to, whereas in Peru, one's status as indigenous is connected to speaking an indigenous language, resulting in higher percentages in the census.

Overall it is clear that the indigenous people of South America are an important part of the continent's demographics. In general South America is a continent of many diversities, because it is not just linguistically and geographically diverse, but it is also diverse in its demographics. As Blouet & Blouet (2015) state, every South American country has a diverse population, since they all have their own heritages and people from different ethnicities. Of the three countries discussed more specifically in this chapter, Peru has the highest percentage and the highest number of indigenous people with 24 % and 7 million people. In Colombia, the indigenous people are located mostly in the interior lowlands, the Oriente, but there are also some smaller ethnic groups located in the Andean areas (Adelaar 2004). The majority of the indigenous people in Ecuador and Peru reside in the inter-Andean valleys, but they also occupy regions in the tropical lowlands (Adelaar 2004). These different aspects of diversities will be used in this thesis as parameters to explain the linguistic differences between languages.

In this chapter I have discussed the linguistic, geographical, demographic and socioeconomic diversities in South America. I will compare languages spoken in the tropical Andes and in the Peruvian Amazon by examining a smaller area within the large region and a smaller set of languages, so I can study the possible linguistic differences in greater detail. This way it is also easier to suggest explanations for the possible structural differences between the languages. The criteria for the research area and the actual language data are discussed more thoroughly in the next chapter.

4 Language data and linguistic parameters

For my study I have chosen a smaller region where languages from several different language families are spoken. In this chapter I will discuss how the research area was selected, and describe the language data used in this thesis. I will use structural features of the languages when measuring their linguistic differences, and these parameters are also discussed in this chapter.

4.1 Language data

The language data used in this analysis was extracted from the World Atlas of Language Structures[1] (henceforth WALS), which is a large database focusing on structural properties of languages. In WALS these properties are called features, and they are the basis for my study. When deciding on the languages and the research area in question, I focused on three overlapping criteria which I thought would help me in getting a relatively representative set of data. First, I wanted to choose a region which would be as diverse as possible in its physical geography. I wanted to include a mountainous area, the vicinity of the sea, but also the vicinity of the Amazonian rainforests. Second, I needed a region where several languages exist, but also that the area includes languages from different language families and also language isolates. The third criteria was the actual language data found in WALS. Since the structural information found in the database is based on descriptive materials, such as reference grammars written by several different linguists, not all languages spoken in the tropical Andes and in the Peruvian Amazon are depicted in WALS, making the selection of the research area a bit more limited.

Moreover, the data in WALS also affected the selection of languages, since I wanted to include languages which have been relatively well described and of which WALS would include a somewhat satisfactory amount of structural features. This meant that even if I found a language in an area, which met all the physical criteria, I needed it to have more than just a few features listed in WALS. After focusing on these three criteria, I managed to locate a region where all these as-

[1] http://wals.info/

Figure 5: Upper scale image of the research area.

pects are met. I was able to find nine languages in that area, which all have enough features and overall fit the aforementioned criteria. Figure 5 represents the research area, which encompasses the entire Ecuador, northern Peru and southern Colombia. There are the mountain range of the Andes, massive rivers and the vicinity of both the sea and the Amazonian rainforest.

It was quite easy to find apt languages from this area, which suited my aforementioned criteria, even though there were languages in this area without any data in WALS. There are of course several other languages spoken in this area and not just the languages used in this thesis, as was mentioned in § 3.2. The nine languages chosen for this study are from six different language families, and three of the languages are categorized as language isolates. The smallest number of features in a single language out of the nine chosen ones is 38 features, while the highest number is 131 features. Figure 6 is a more detailed depiction of the chosen languages and their locations in the area. The nine languages are: Camsá (C), Yagua (Y), Murui Huitoto (M), Jebero (J), Resígaro (R), Epena (E), Páez (P), Awa Pit (A) and Shuar (S).

The locations for the languages are adapted from the South American Phonological Inventory Database[2] and WALS. I of course understand that the language areas are much wider than just the pinpointed regions for every language. However, in this thesis the most important aspect is the approximate whereabouts of these languages, which according to WALS and the South American

[2] `http://linguistics.berkeley.edu/~saphon/en/`

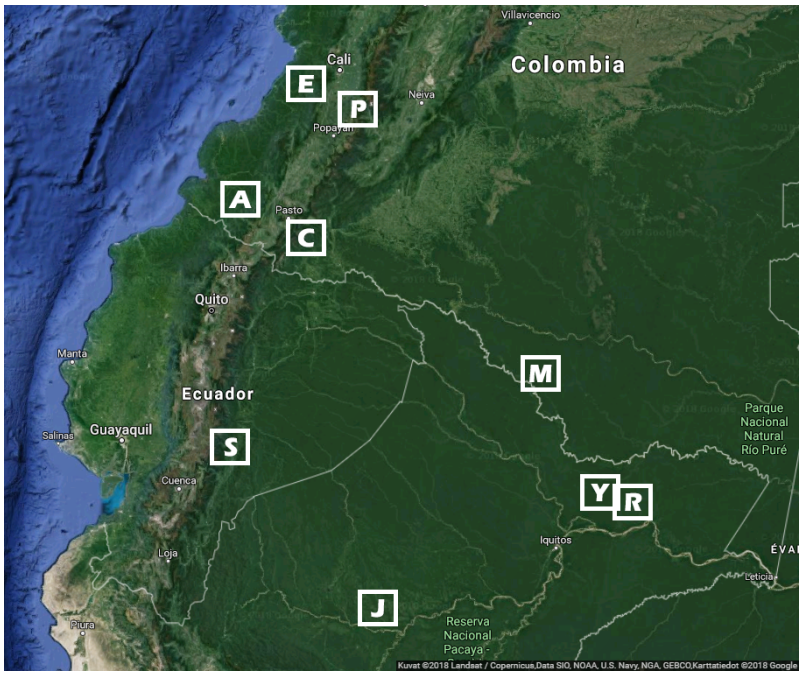

Figure 6: The research area and the languages studied in this thesis.

Phonological Inventory Database would be around the marked areas. Next I will briefly describe these nine languages, giving a somewhat general summary of their prominent characteristics. The following descriptions portray the typological profiles of these languages, which is the basis for the analysis conducted in this thesis.

Camsá is a language isolate, spoken in Colombia. According to the 2008 census, Camsá has about 4,000 speakers and its status is developing, which means that the language is used to write. Typologically it does not have a dominant order for subject, object and verb, and neither for object and verb nor for adjective and noun, which is definitely interesting. Camsá has 38 features listed in WALS, which is the smallest number of features for a language in this study.

Yagua is a language isolate spoken in northeast Peru. According to the 2000 census, it has around 5,700 speakers, and it is categorized as a threatened language. Typologically Yagua is a VSO language, making it the only VSO language in this study. Additionally it has a large amount of postpositions, a simple tone system and also a dual number. In WALS Yagua has 127 listed features.

Murui Huitoto belongs to the Witotoan language family, and it is spoken in Colombia and Peru. According to the 2002 census, Murui Huitoto has 7,800 speakers, but its vitality status is still considered to be threatened. Typologically Murui Huitoto is an SOV language, and it is highly synthetic and has a large classifier system. Murui Huitoto has 48 features listed in WALS.

Jebero is a Cahuapanan language spoken in Peru. According to the 2006 census, Jebero has 2,500 speakers, and it is categorized as shifting language, which means that another language is used more than Jebero by its speakers. Typologically Jebero, like Camsá, does not have any dominant order for any of the main constituents. It has 46 features listed in WALS.

Resígaro is an Arawakan language spoken in Colombia and Peru. In 1976 Resígaro was recorded to have only 14 speakers, and other studies claim that in 2017 there was only one person who could speak Resígaro. Hence its vitaly status is nearly extinct, which is the lowest status in this study. Typologically Resígaro is an SOV language, and it has a simple tone system. Resígaro has 59 features listed in WALS.

Epena is a Chocoan language spoken in Colombia and Ecuador. According to the 2004 census, it has 3,500 speakers. Unlike the other languages used in this study, Epena's language vitality status is in the upper end of the vitality scale, being educational, which means that it is used in education. Typologically Epena is an SOV language, and it has no nasals. Epena has 121 features listed in WALS.

Páez is a language isolate spoken in Colombia. According to the 2007 census it has 40,000 speakers, which is quite a lot compared to the other languages used in this study. Despite its number of speakers, it is still categorized as a threatened language. Typologically it is an SOV language, and it has a complex syllable structure. Páez has 60 features listed in WALS.

Awa Pit is a Barbacoan language spoken in Colombia and Ecuador. According to the 2008 census, it has 13,000 speakers, and its language vitality status is threatened. Typologically Awa Pit is an SOV language, and it does not differentiate any remoteness distinctions in the past tense, and it has 10 or more cases. In WALS it has 131 features listed, which is the largest number of features for a language in this study.

Shuar is a Jivaroan language spoken in Ecuador. According to the 2007 census Shuar has 35,000 speakers, and its vitality status is developing. Typologically it is an SOV language, and it is a strongly suffixing language. In WALS Shuar has 51 listed features.

To summarize, the nine languages I have chosen clearly represent a set of languages with a variety of different characteristics. These factors are shown in Table 4. By choosing these nine languages to be used in this thesis, I managed to include languages from a wide range of different language families and also language isolates. The number of speakers also varies, ranging from 14 all the way to 40,000. Unfortunately I had to rely on the available census data when accounting for the number of speakers for each language, resulting in numbers which are at minimum ten years old. The data for Resígaro is not even from the 21st century, so counting on that data is somewhat questionable, but due to the lack of recent census data, it is the number I have to trust.

Table 4: Language data used in this thesis.

	Genealogy	Speakers	Area	Status	Constituent Order	Features in WALS
Camsá	Isolate	4,000	Colombia	Developing	No dominant order	38
Yagua	Isolate	5,700	Peru	Threatened	VSO	127
Murui Huitoto	Witotoan	7,800	Colombia, Peru	Threatened	SOV	48
Jebero	Cahuapanan	2,500	Peru	Shifting	No dominant order	46
Resígaro	Arawakan	14	Colombia, Peru	Nearly extinct	SOV	59
Epena	Chocoan	3,500	Colombia, Ecuador	Educational	SOV	121
Páez	Isolate	40,000	Colombia	Threatened	SOV	60
Awa Pit	Barbacoan	13,000	Colombia, Ecuador	Threatened	SOV	131
Shuar	Jivaroan	35,000	Ecuador	Developing	SOV	51

The age of the census data makes it hard to say anything about the more recent status of these languages, whether some of them have already died or possibly survived their threatened status, or even managed to slow down the decline process, so a more up to date census data would have been ideal. As expected, almost all of these languages are either threatened or shifting towards very serious endangerment, ranging from level 4 (Educational) to level 8b (Nearly extinct) on SIL's language vitality scale[3]. Out of the nine languages, there are two, Páez and Awa Pit, which are recognized as threatened languages, despite their numbers of speakers being more than Epena, which has only 3,500 speakers and is still used in education. This is a great example of how people should not declare anything about languages' vitality status based solely on the number of its speakers, since languages are very complex entities and more than just the number of speakers.

Typologically there is not that much variance in the ordering of the constituents. SOV is the most prominent order, with the exception of Yagua, which is a VSO language, and two languages, Camsá and Jebero, which do not have a dominant order for the constituents. The number of features listed in WALS is the most essential aspect for this study, since these features are used to conduct the statistical analysis, and the differences between these languages are based on their respective feature values. It would not be wise to include all of the features for every language, since that would possibly result in a skewed representation of their differences. Hence I needed to select a suitable set of features, which all of these languages would have. These parameters are discussed in the next section.

[3] `https://www.sil.org/about/endangered-languages/language-vitality`

4.2 Linguistic parameters for the study

The data in WALS conforms to the statements made in § 2.2.1, where I mention how grammatical elements are the most reliable when comparing differences between languages. Before I can analyze these languages statistically, it is important to determine the relevant structural parameters for each language from their respective lists of features represented in WALS. I selected as many appropriate features as I might find which all the languages would have in common. As a result, I found six features represented in all nine languages, which are as follows:

(1) Negative morphemes
(2) Order of subject, object and verb
(3) Coding of nominal plurality
(4) Prefixing vs. suffixing in inflectional morphology
(5) Position of tense-aspect affixes
(6) Consonant-vowel ratio

Even though the data represented in WALS varies quite a lot between languages, that being its deficiency, I believe these six features cover different structural properties quite well. All the

Table 5: Features and their values.

Features		
Negative morphemes (neg)	*Order of subject, object and verb (svo)*	*Coding of nominal plurality (nominal)*
(1) Negative affix (2) Negative particle (3) Negative auxiliary verb (4) Negative word, unclear if verb or particle (5) Variation between negative word and affix (6) Double negation	(1) SOV (2) SVO (3) VSO (4) VOS (5) OVS (6) OSV (7) Lacking a dominant word order	(1) Plural prefix (2) Plural suffix (3) Plural stem change (4) Plural tone (5) Plural by complete reduplication of stem (6) Morphological plural with no method primary (7) Plural word (8) Plural clitic (9) No plural
Prefixing vs. suffixing in inflectional morphology (presuf)	*Position of tense-aspect affixes (tense)*	*Consonant-vowel ratio (conso)*
(1) Little or no inflectional morphology (2) Predominantly suffixing (3) Moderate preference for suffixing (4) Approximately equal amounts of suffixing and prefixing (5) Moderate preference for prefixing (6) Predominantly prefixing	(1) Tense-aspect prefixes (2) Tense-aspect suffixes (3) Tense-aspect tone (4) Combination of tense-aspect strategies with none primary (5) No tense-aspect inflection	(1) Low (2) Moderately low (3) Average (4) Moderately high (5) High

features listed in WALS have a set of different values, ranging from 2 to 28 values. For example, the feature 3A in WALS, Consonant-vowel ratio, has 5 different values: low, moderately low, average, moderately high and high. Table 5 contains the values for the features chosen for this study. It is clearly visible from the table that these features all have a wide range of values.

Essentially all the languages I am using in this study have a certain value for each of the features listed in the table above. However, using these feature values in a statistical software and comparing them would probably not render any kind of proper results, since the software needs to calculate the distances for each feature. There is not much to calculate in the distance between the concepts of "SOV" and "negative affix" – that would be rather redundant. Luckily WALS also provides numerical data on these feature values, which means that each value has been assigned a number. For example, Camsá has negative affixes, which corresponds to the numerical value of 1 in this set of values.

Table 6: Numerical values for the features.

	neg	svo	nominal	presuf	tense	conso
Camsá	1	7	2	5	1	3
Yagua	2	7	2	2	2	1
Murui Huitoto	1	1	2	2	2	2
Jebero	1	7	2	2	2	4
Resígaro	2	1	2	4	2	4
Epena	1	1	8	2	2	2
Páez	1	1	2	2	2	5
Awa Pit	6	1	2	2	2	3
Shuar	1	1	2	2	2	3

Table 6 presents all the numerical values for each language and for feature, and with these values it is easier to measure the linguistic differences of these languages. Some features have quite a lot of variety in their values, such as "conso" (Consonant-vowel ratio), since all the numbers from 1 to 5 are represented, but some feature values are almost in unison, such as "nominal" (Coding of nominal plurality) with just two different values. The feature values represented in Table 6 will be used as the language data in order to measure these values statistically. The measurements will showcase the linguistic differences between the languages, and these differences will eventually be visualized in a map. Visualizing the results of the statistical measurements will reveal underlying relations in the data, which would otherwise remain unnoticed just by looking at tables with numerical values, such as Table 6. The statistical method used in this thesis is described thoroughly in the next chapter, followed by the results of the analysis.

5 Statistical method and results

The research method used in this thesis is quantitative, since I am employing a multidimensional scaling method in comparing a set of languages in order to measure their linguistic differences. In this chapter I will describe how a statistical software R is used to conduct the analysis. In the end I will present the results of the analysis.

5.1 Description of R[1]

Since I am using R for calculating the data and displaying it graphically, it is quite relevant to briefly discuss what R is all about. The R interface is basically a software for statistical computing and graphics, but its roots lie in the implementation of the S language, which was developed in the United States in the 1980s (Venables & Smith 2009). R has a wide range of different applications, but first and foremost it is a method of interactive data analysis (Venables & Smith 2009). So to put it shortly, the R interface is a programming environment (Gries 2013). The beauty of R is its freedom, because it is an open-source software, so it is possible to create new methods and tools for analysis, or to use methods which other users have developed (Oksanen 2003).

Over the years R has grown to be used widely throughout different disciplines for its possibilities in statistical computing and graphics. Statistical analysis is not a new phenomenon, and linguists should explore its possibilities, since any kind of observed data in a study most likely demands statistical treatment (Gries 2013). By this type of treatment linguists can analyze numerical data and also interpret it. Some linguists have been rejecting quantitative research, even though qualitative and quantitative research go hand in hand, since the idea is not to work against the other method, but to work in unison in order to get the best results (Gries 2013).

R has a command line based user interface, where the commands are written after the command prompt (>) one at a time (Baayen 2008). Technically R is an expression language, and it has a very simple syntax, since the basic commands consist of expressions or assignments. There are several types of data which can be analyzed in R, for instance vectors, matrices, data frames and lists, and in general there are hundreds of different functions and ways of manipulating data.

[1] I would like to thank Dr. Kaius Sinnemäki for his invaluable advice regarding statistical methods and R.

Even though I must explain and show what I am doing with R, and even though the method used in this study is a complex one, I want to focus on making my explanations as understandable as possible, since I want other linguists as well to fully understand what I am doing.[2]

5.2 Analyzing linguistic data

The easiest way of comparing a set of languages in order to measure their linguistic differences is to use statistical methods. There are dozens of ways of analyzing language data statistically, but luckily there is a family of clustering methods suitable for studying and measuring linguistic differences. These methods are used to examine relations between many variables. Basically, the relevant data sets are compiled as matrices, so that the matrices represent the observed data, where the rows list the observations and the columns specify the different variables (Baayen 2008). This type of data with multiple vectors is called multivariate data, and Baayen lists five different clustering methods for studying multivariate data:

(1) Principal component analysis
(2) Factor analysis
(3) Correspondence analysis
(4) Multidimensional scaling
(5) Cluster adnalysis

Of these five methods the first two are used to analyze tables with measurements, the third is used to analyze tables with counts, while the last two methods are used to analyze tables with distances. McEnery & Wilson (2001: 88) state how all of these methods have a common goal, which is to "summarize a large set of variables in terms of a smaller set on the basis of statistical similarities between the original variables, whilst at the same time losing the minimal amount of information about their differences". To put it shortly, the goal is to find structure in the data by grouping the observations. The task is then to provide meaningful interpretations of these findings.

Since I want to measure linguistic differences, I need a method to analyze distance tables. Both multidimensional scaling and cluster analysis would suffice, but I chose multidimensional scaling as my research method, because it represents linguistic differences as spatial distances preferably on a two-dimensional xy-plane (Schmidtke-Bode & Hetterle 2008). This spatial model summarizes the differences in the data, which can then be interpreted. The multidimensional scaling method is further discussed in the next section.

[2] For those who want to deepen their knowledge on R's commands and functions, I recommend exploring RDocumentation (`https://www.rdocumentation.org/`), which is a database for all the packages used in R. A fast track to learning R might be possible by browsing different forums related to R's commands and functions, if RDocumentation seems too complicated.

5.3 Multidimensional scaling

Multidimensional scaling (henceforth MDS) is defined as "a technique for the analysis of similarity or dissimilarity data on a set of objects" (Borg & Groenen 2005: vii). More specifically, it refers to the way of visualizing underlying relational structures within the data (Hout et al. 2013). With MDS it would be possible to analyze any kind of distance matrix, which could represent any kind of similarities between any kind of features. Possible features could be numerical data, class variables or different quantities. The analysis is done by computing the MDS algorithm, which calculates the distance of rows in a data matrix and then models the distances among points in a geometric space (Borg & Groenen 2005: vii). The geometric space, i.e. a map, is the outcome of MDS, which depicts the relationships between the objects, where similar objects are located in close vicinity to one another and different objects are located further apart (Hout et al. 2013). The purpose of this method is to detect meaningful relationships in the data, so it would be possible to explain the differences, i.e. the distances, between the objects (StatSoft 2013). The easiest way of understanding the MDS algorithm and its results is by a question: "How similar is A to B?". The MDS map shows how similar A is to B, and if they are different, it shows how different they actually are.

The roots of MDS are in behavioral sciences, but nowadays it is used in many disciplines which can benefit from statistical techniques. One of these disciplines is linguistics, where MDS can be used for example in semantic typology for visualizing semantic maps (Croft & Poole 2008). MDS is also useful for comparing typological profiles of languages, which is exactly what I am focusing on in this thesis. In relation to my own thesis, I want to find out how different a group of South American languages is. The idea is to have a representative data set of the group of languages and their features, and then to analyze the features by using the MDS method.

There are two different ways of analyzing data, either by metric or non-metric MDS. Metric MDS is done with the algorithm `cdmscale` and non-metric MDS is done with `isoMDS` (Schmidtke-Bode & Hetterle 2008). These two algorithms differ in they way they interpret the data, because `cdmscale` views the calculated distances just as numbers, whereas `isoMDS` sees them as ordinal numbers (Everitt & Hothorn 2006). Non-metric MDS is also better for a data set which has categorical data. Because the data for the feature `conso` is categorical, the use of the non-metric algorithm `isoMDS` is justified. Additionally, the non-metric MDS represents the differences between objects as closely as possible, and it can be applied to any kind of distance matrix (Clarke 1993). Yet another reason for using `isoMDS` is its usefulness in calculating a goodness-of-fit measure, which indicates how many dimensions MDS should use to show the differences found in the data in the best possible way.

Performing the MDS method requires three things: the modification of the data, the ordering of the data, and the creation of a data matrix. First, the feature values represented in Table 6

(see § 4.2) need to be modified into factors, which means that the feature values will be stored as integers (1, 2, 3, . . .). This way R can treat the data as having predefined values, and not as strings, i.e. characters. Second, the numerical values of the feature conso (Consonant-vowel ratio) need to be ordered accordingly, because that particular feature represents categorical values (low-high), which can be put in an order. Third, the modified data needs to be converted into a distance matrix by using the function `daisy()` from the R package *cluster* (Maechler et al. 2017), which calculates the differences between the variables in the data. Because the language data has class variables, `daisy` is the optimal function for the calculations. Additionally, the occurrence of class variables implies that `daisy`'s algorithm should use Gower's coefficient as its metric, because it measures how different variables are (Gower 1971). The result is a distance matrix, which is used as the argument of the non-metric MDS function `isoMDS()` from the R package *MASS* (Venables & Ripley 2002). Finally, the values calculated for a suitable number of dimensions will be mapped into a scatter plot by using the function `plot()`, which results in a visual representation of the differences found in the language data.

5.4 Results

In this section I will present the results of the MDS analysis. The description of the results intends to be as encompassing as possible, but still maintaining a certain clarity. The analysis starts by changing the feature values in Table 6 into factors with a simple `for` command:

```
> for(i in 1:6){
      amer[,i] = as.factor(amer[,i])
}
> amer
              neg svo nominal presuf tense conso
Camsá           1   7       2      5     1     3
Yagua           2   7       2      2     2     1
Murui Huitoto   2   1       2      2     2     2
Jebero          1   7       2      2     2     4
Resígaro        2   1       2      4     2     4
Epena           1   1       8      2     2     2
Páez            1   1       2      2     2     5
Awa Pit         6   1       2      2     2     3
Shuar           1   1       2      2     2     3
```

Here the command `as.factor()` modified the data type found in the variable `amer`. The result is a table of factors, which looks identical to Table 6, so this command has not changed anything externally, instead it has just changed the data type into categories. Because `conso` represents

categorical values, they cannot be ordered randomly, but they need to be ordered according to their levels. Luckily, there is a simple command called `ordered`, which fixes the order of the levels:

```
> amer$conso = ordered(amer$conso, levels = 1:5)
> amer$conso
[1] 3 1 2 4 4 2 5 3 3
Levels: 1 < 2 < 3 < 4 < 5
```

The next step is to actually calculate the distances between the observations in the data set. As mentioned in the previous chapter, this is done by the command `daisy()`, as follows:

```
> amer.dist = daisy(amer, metric = "gower")
> amer.dist
Dissimilarities :
                  Camsá      Yagua Murui Huitoto     Jebero   Resigaro
Yagua         0.58333333
Murui Huitoto 0.70833333 0.20833333
Jebero        0.37500000 0.29166667    0.41666667
Resigaro      0.70833333 0.45833333    0.25000000 0.50000000
Epena         0.70833333 0.54166667    0.33333333 0.41666667 0.58333333
Páez          0.58333333 0.50000000    0.29166667 0.20833333 0.37500000
Awa Pit       0.66666667 0.41666667    0.20833333 0.37500000 0.37500000
Shuar         0.50000000 0.41666667    0.20833333 0.20833333 0.37500000
                   Epena       Páez    Awa Pit
Yagua
Murui Huitoto
Jebero
Resigaro
Epena
Páez          0.29166667
Awa Pit       0.37500000 0.25000000
Shuar         0.20833333 0.08333333 0.16666667
```

The result is a distance matrix, which is assigned to the variable `amer.dist`. Before I can use the non-metric MDS method, I need to verify what is the most suitable number of dimensions to express the data well enough. This is done by calculating the goodness-of-fit measure, which depicts the stress value of the data. The stress value is based on deciding how many dimensions are appropriate for the analysis by comparing the actual distances and their predicted values (Hout et al. 2013; StatSoft 2013).

First I need to calculate the MDS values from the distance matrix for all the different dimensions, and then use those values for pointing out the "elbow" in the values. This "elbow" is the point in the results where the line starts to be skewed, meaning that adding more dimensions than the "elbow dimension" will not show anything significant with respect to the data (Johnson 2008). Calculating the MDS values from the distance matrix by the command `isoMDS()` is quite simple, and this is done for several dimensions as follows:

```
> amer.k.1 = isoMDS(amer.dist, k = 1)
initial  value 35.314642
iter   5 value 27.794205
final  value 26.972894
converged
> amer.k.2 = isoMDS(amer.dist, k = 2)
initial  value 13.375152
iter   5 value 9.859714
final  value 9.709941
converged
> amer.k.3 = isoMDS(amer.dist, k = 3)
initial  value 7.883363
iter   5 value 4.633528
iter  10 value 4.146508
iter  15 value 3.940456
iter  20 value 3.846236
iter  25 value 3.765291
iter  25 value 3.763688
iter  25 value 3.763224
final  value 3.763224
converged
> amer.k.4 = isoMDS(amer.dist, k = 4)
initial  value 4.748101
iter   5 value 1.943087
iter  10 value 1.543621
iter  15 value 0.900106
iter  20 value 0.609273
iter  25 value 0.538911
iter  30 value 0.350690
iter  35 value 0.278007
iter  40 value 0.188017
```

```
iter  45 value 0.175963
iter  50 value 0.163608
final  value 0.163608
```

Here `isoMDS()` uses the distance matrix to calculate the MDS values, and the additional argument `k` means the respective number of dimensions in each case. R prints the results straight away, and it is already visible where the stress value drops considerably.

Since it is also useful to visualize the "elbow", I need to plot the stress values on a graph by using the final values for each dimension:

```
> amer.stress = c(amer.k.1$stress, amer.k.2$stress, amer.k.3$stress,
+ amer.k.4$stress)
> dim = 1:4
> plot(dim, amer.stress, lines(dim, amer.stress, type="l"),
+ main="Goodness-of-fit Measure: Stress Values",
+ xlab="Number of Dimensions", ylab="Stress")
```

The stress values are assigned to a variable `amer.stress` and the number of dimensions is assigned to `dim`. The function `plot()` literally plots the stress values onto an xy-plane, where x represents the number of dimensions and y the stress values. From Figure 7 it is clearly visible that the "elbow" in the values is located between the second and the third dimension. According to Johnson (2008), stress values should not be over 20, but preferably under 10. My goodness-of-fit measure clearly shows how a two-dimensional representation of the data is adequate and can be used in this study.

With this in mind I will proceed in mapping the language distances onto a two-dimensional xy-plane. Because I have already calculated the MDS values for two dimensions, the correct values for the map are already assigned to the variable `amer.k.2`, which is easily checked by just calling it:

```
> amer.k.2
$points
                     [,1]         [,2]
Camsá          0.50596839  0.005649622
Yagua          0.06654686 -0.238437508
Murui Huitoto -0.12953230 -0.103298062
Jebero         0.15898506  0.028489530
Resigaro      -0.23006765 -0.247203639
Epena         -0.10733209  0.307913915
Páez          -0.04368438  0.119942021
Awa Pit       -0.20745221  0.039907329
Shuar         -0.01343167  0.087036792
```

Figure 7: Goodness-of-fit measure of the data.

The next step is to plot these values onto a two-dimensional plane, which is done with the same `plot()` function as above:

```
>  plot(amer.k.2$points[,2:1], type="n", xlab = "Dimension 1",
+ ylab = "Dimension 2",
+ main = "Multidimensional Scaling of Linguistic Differences",
+ xlim = c(-0.3,0.35))
> text(amer.k.2$points[,2:1], as.character(rownames(amer)), cex=1.0)
```

This function looks quite confusing, but its arguments are pretty straightforward. First I used the points found in `amer.k.2`, and then I named both axes and the entire table, respectively, and then I determined the range for the x-axis. The names of the languages are easily plotted into the map by the function `text()`, where the only adjustable argument, `cex`, is the font size. The result is an MDS map, which represents the distances among the languages used in this study.

Figure 8 is the result of this language distance analysis, which shows how different the nine languages are structurally. As is clearly visible, these languages are scattered somewhat widely

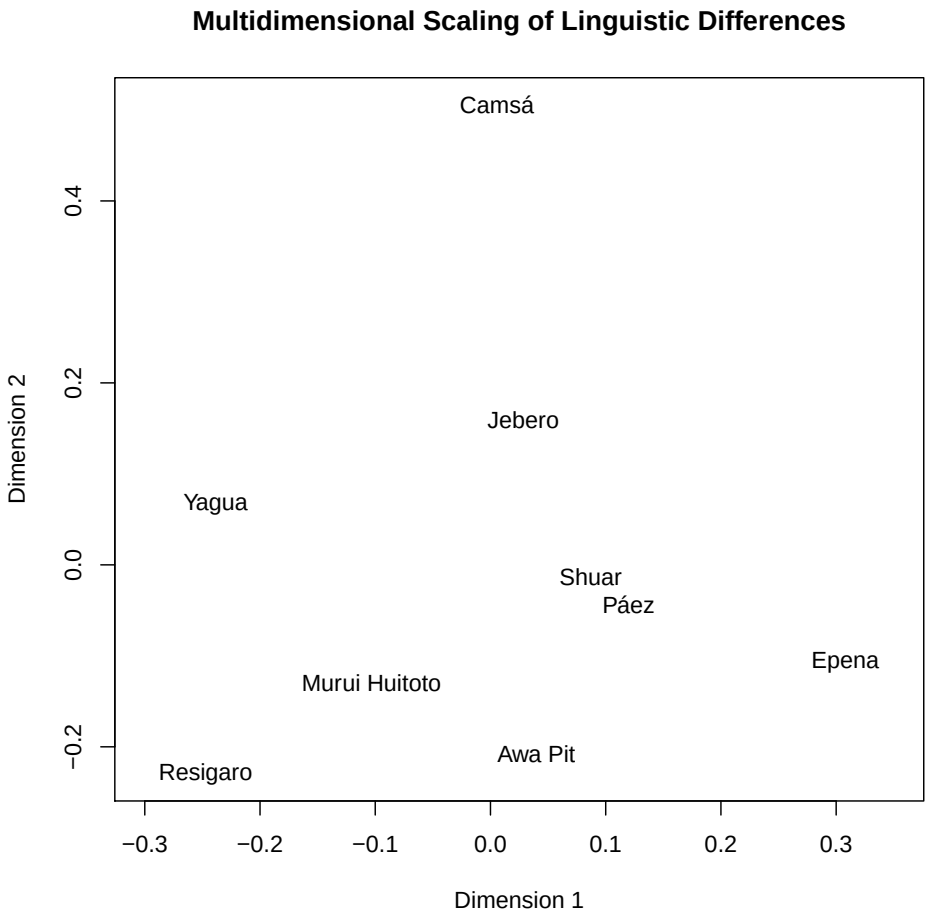

Figure 8: Multidimensional scaling of the data.

on the plane, but nevertheless some languages are closer to others, such as Shuar and Páez, and some, for example Camsá, are located almost in isolation from the other languages. Of these nine languages the two isolates, Yagua and Camsá seem to differ the most from the other languages, while the third isolate Páez does not show significant linguistic differences. Most of the other languages belonging to a specific language family are located in the middle area of the plane, showing smaller differences.

The quantitative part of this thesis focused on measuring linguistic differences, and this chapter has described the method in action quite meticulously. The result is a linguistic distance map, which represents the chosen languages and their differences as spatial distances in a geometric space. The qualitative part of this thesis occurs in the next chapter, in which I will interpret these differences and furthermore, I will suggest possible explanations for them.

6 Discussion

The results seen in Figure 8 showcase the linguistic differences between the languages. Interestingly, there are some intriguing differences among the languages, such as the distance between the two isolate languages compared to the other languages, and the close proximity of some languages in the distance map compared to their actual geographical distances. For example, Shuar is spoken near the border of Peru in Ecuador, while Páez is spoken in the Andean Colombia. Shuar belongs to the Jivaroan language family and Páez is a language isolate. Nevertheless, based on the calculations, they are both situated almost in the same location on the MDS map, which means that their linguistic differences are quite minimal. The situation is opposite with Resígaro and Yagua, an Arawakan language and another language isolate. Their close geographical vicinity is not visible on the MDS map, on which they are quite far away from one another, which illustrates linguistic differences between them. Yet another interesting difference is found between Camsá, an isolate, and Awa Pit, a Barbacoan language, which are both spoken in Colombia, near the border of Ecuador. Awa Pit is spoken on the coastal side of the Andes, while Camsá is spoken on the eastern side. This close proximity does not seem to have influenced these languages structurally, because on the MDS map their distance from one another is quite significant.

As is clearly visible from these few examples, many interesting details can be found on the MDS map. Just listing these findings would lack a perspective I believe linguists should embrace more often – the task of explaining the hows and whys of these findings. Why are two totally different languages, located in two totally different areas in South America almost exactly similar when it comes to their calculated structural differences? How come two different languages found in close proximity to one another are actually structurally quite different? These are very interesting questions that require a more thorough analysis. In this chapter I will analyze the linguistic differences between the languages and offer possible explanations for them. Additionally, I will also explain diachronically how and why the overall linguistic diversity in South America and in the research area has changed, basing my explanations on Dixon's punctuated equilibrium model.

6.1 Explaining linguistic differences

The theoretical approaches presented in Chapter 2 offer ways of explaining the linguistic differences visible on the MDS map. I have divided these explanations into two categories: non-linguistic and linguistic explanations. The non-linguistic explanations are based on the ecological view on language, focusing on the geographical and socioeconomic factors, whereas the contact-induced language change is used as an linguistic explanation. The complicated relationship between language and its environment reflects how people commonly identify themselves as belonging to a certain society and to certain groups by the use of a specific language. Hence the selection of the two distinct approaches. I feel these explanations might make us look at these languages, or comparative linguistics in general, in a totally new light, because including both qualitative and quantitative aspects in language studies is beneficial.

6.1.1 Non-linguistic explanations

The geographical explanations presented in this section are based on the elements in the natural environment, such as river systems and mountains, whereas the socioeconomic explanations are connected to the different social, economical, demographic and political aspects of the human environment. These explanations are used consecutively to analyze the linguistic differences found on the MDS map.

6.1.1.1 Geographical explanations

The research area comprises of the areas in the tropical Andes and in the Peruvian Amazon, encompassing southern Colombia, Ecuador, and northern Peru. All three countries can be described geographically as consisting of three distinct areas: the flat coastal area west of the Andes, the central highland area of the Andes, and the eastern lowland plains, as described in § 3.3. These three recognized areas each have their own specific climates and vegetations. The coastal areas between the Andes and the Pacific are mostly grass savannas, semideserts or tropical rainforests, depending on the latitude. This variation is also visible in the coastal climate, since it is tropical, but depending on the specific location, the area has either dry winters or sufficient rainfall in all months. Understandably, the Andes are mostly alpine tundra or alpine forests experiencing a highland climate. The lowland plains on the other hand are characterized by equatorial tropical evergreen rainforests, warm temperatures and year-round rainfall. In addition to the Andes, also the number of rivers is a dominant feature. The world's largest river by discharge volume, the Amazon river, begins at the junction of two Peruvian rivers, and it covers almost 5 million square miles (Blouet & Blouet 2015). So it is safe to say that the rivers of the Amazonian lowlands also dominate the scenery.

This description serves as the basis for explaining how geographical parameters can influence the languages used in this study. First I need to compare the geographical and linguistic distances, which is done by comparing a topographic map and the MDS map. This way it would be easier to understand the geographical locations of the languages, and how they correlate with the linguistic distances of the languages. Figure 9 represents this comparison between these two distances, where each letter on the left-hand side map represents the languages used in this study. Right

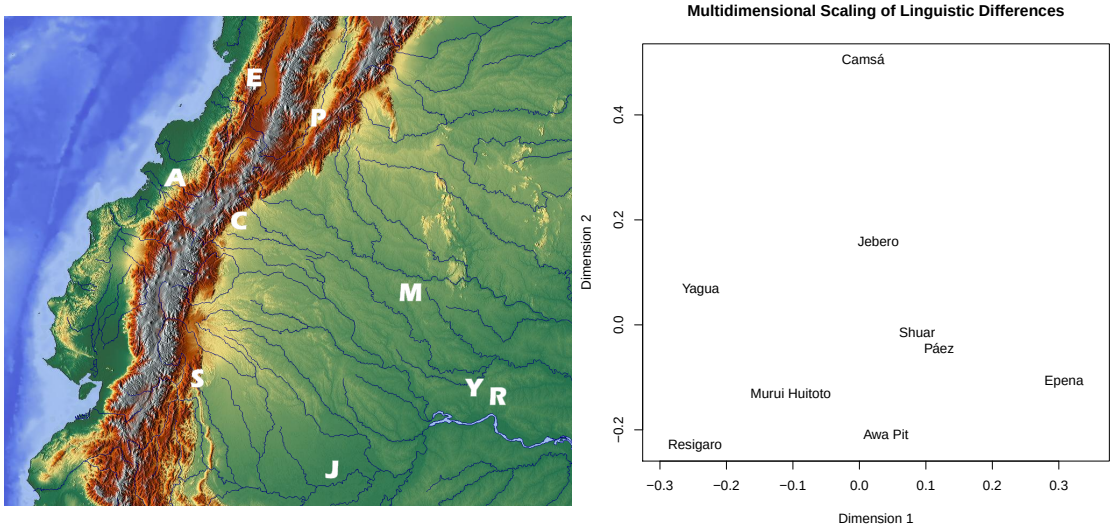

Figure 9: The comparison between geographical and linguistic distances.

away it looks like the geographic distances are not equivalent to the linguistic distances of these languages. Languages which are spoken quite close to one another physically are not as close, i.e. similar, linguistically, and vice versa; languages located in isolation on the topographic map are actually not so distant from the other languages structurally. For example, both Epena & Páez and Awa Pit & Camsá are spoken in the same regions, in the highlands of the Andes in Colombia. These four languages are quite different linguistically based on the MDS map, since they are located far from each other, each in their own respective location on the map: Epena on the right, Páez almost on the middle, Camsá above and Awa Pit below.

Based on this example it is clear that the geographic distances do not necessarily correspond to the linguistic distances. The most intriguing question to ponder is why that might be the case. One clear and rather dominant aspect is the presence of the Andes, especially in relation to the four aforementioned close-proximity languages. First, understanding the physical features of the mountain range is crucial in understanding how these languages have stayed structurally different. The Andes have enabled the indigenous groups to maintain their own societies in isolation, since the Europeans deemed the colonization of the rugged areas as too difficult and unprofitable (Adelaar 2004). As described above, the variation in climate and vegetation is also a very distinct

feature of the area. This variety around and within the Andes has resulted in the possibility of each ethnic group residing in different ecosystems (Adelaar 2004), which means that groups can prosper on their own within their own habitats.

For example, in Colombia the Andes are bisected by river valleys and large forests, which has been pivotal for the survival of the ethnic groups (Adelaar 2004). Blouet & Blouet (2015) also state how the Andes are densely populated despite the rugged physical appearance of its high peaks and steep slopes – even at high altitudes it is possible to cultivate land. To sum up, these features have enabled the indigenous groups speaking Awa Pit, Epena, Páez and Camsá to survive in isolation within the Andes. The isolation has maintained the lack of contact to other groups, which might help explain the significant structural differences found on the MDS map regarding these languages. They are all different in size, since, for example Páez has ten times more speakers than Camsá (see Table 4). Geographical aspects of the research area have indeed shed light on the relationship between these four Andean languages and on their structural differences.

The four aforementioned languages demonstrate differences despite their close geographical proximity, but even more interesting is the distances between Yagua and Resígaro, which are both spoken in northeastern Peru in the vicinity of the Amazon river. Even though they are spoken in the same region, they illustrate significant linguistic differences based on the distances on the MDS map. The difference might be explained by their genealogy, since Yagua is a language isolate and Resígaro is an Arawakan language, but matters are not always so straightforward. Here I cannot look for geographical explanations from the presence of the Andes, because these languages are both spoken in the Amazonian lowlands, quite far from the Andes. Instead, I must rely on the presence of the massive river system of the Amazon. The Yaguan people are scattered across the lowlands in several small villages, and the distance between some villages might be quite substantial (Payne 1985). The massive river system has enabled smaller groups of Yaguans to migrate into remote locations, where they can prosper in seclusion without being in contact with other ethnicities. This seclusion has enabled the language to maintain its structural diversity among other indigenous languages, explaining why the calculated difference between Yagua and Resígaro is so evident.

Another baffling result of the MDS is the slight structural difference between Shuar and Páez, of which Shuar belongs to the Jivaroan language family and Páez is an isolate, as was already mentioned above. As it is clearly visible from Figure 9, their geographical distance is apparent, since Shuar is spoken in southern Ecuador near the border of Peru, while Páez is spoken in the Andean Columbia. If I were to seek explanations for their small structural differences from geographical parameters, I would suggest that the Shuar and the Páez people have utilized the river systems, especially the river valleys which cut through the Andes in Colombia and provide access from the Amazon basin all the way to the Pacific. So despite their geographical distance, the landscape of South America might have facilitated the communication between the Shuar and

the Páez people, resulting in the structural similarities. However, I cannot rely on this suggestion, since the result of the MDS measure might just be a coincidence due to the number of features used in this study being only six. Looking for more solid justification for their structural similarity would require a more elaborate analysis of their structural features, which is unfortunately not the scope of this study.

By looking at geographical parameters I have managed to suggest possible explanations for the linguistic differences found on the MDS map. For some languages it was quite obvious how their geographic location has influenced the development and the preservation of the language. It is also clear that the relationship between languages and their ecological environment is a complex one.

6.1.1.2 Socioeconomic explanations

Another way of explaining the linguistic differences found on the MDS map can be found by using socioeconomic parameters. The rise of hegemonies has affected the South American Indian languages, since communities are shifting from their indigenous languages to other languages, mainly to Spanish and Portuguese. However, it is important to keep in mind that based on the numbers of speakers and the vitality status of the research languages, the language shift in South America has been ongoing for a while. For most ethnic groups the reason behind the inevitable language shift is the lack of active language maintenance, which goes hand in hand with the ignorant attitude of the ethnic groups, since they do not seem to recognize the benefits of preserving their own languages. For some the shift has been quite rapid, since young people want to get on with the urban developing societies, but some have managed to avoid the shift and maintain their ethnic language by living in isolation, usually in rural conditions. Holmes (2013) identifies four different factors influencing and causing language shift: economic, social, political and demographic factors. I will use these socioeconomic parameters as the basis for the possible explanations for the attested linguistic differences.

In the previous section I mentioned how Yagua and Resígaro demonstrate structural differences despite their close geographical proximity to one another. The case of Yagua and Resígaro is also interesting from a socioeconomic perspective. At the time of writing, Resígaro is probably already an extinct Arawakan language, since the only available census data is from 1976, when Resígaro had 14 speakers. Understandably, a highly endangered language with low prestige will not affect a language isolate no matter how close, but at some point in time Resígaro has been a thriving language with its own ethnic groups residing in the Amazonian lowlands. Nevertheless, it seems that these two ethnic groups have not had any economic or social contacts outside their own villages. The Yaguans commonly practice cultivation and fishing, but some groups of Yaguans have remained hunter-gatherers (Payne 1985). The Yaguan villages are quite remote and located far from each other, which is due to the natural resource exploitation, when the Euro-

peans extracted rubber from the area at the turn of the 20th century. This exploitation resulted in Yaguans fleeing to even more remote areas than before, which meant that people belonging to the same ethnicity could not properly contact each other. Because the groups are dispersed, Yaguans rarely marry outside their own tribe.

These aspects have only strengthened the isolation of this ethnic group from others, and I believe the lack of social contacts outside the villages has enabled the Yaguan language to remain structurally diverse as was shown on the MDS map. This explains why the linguistic differences between Resígaro and Yagua are quite evident, and it also supports the language status of Yagua as a language isolate. However, the situation might be different in the future. Due to the seclusion of Yaguan villages, people have started to show growing interest in the socioeconomic advantages gained from speaking Spanish, because they rarely have any contact with other Yaguans. Time will tell if Yagua will remain a structurally different secluded language isolate or whether it will also start to show signs of language shift.

The four Andean Colombian languages of Epena, Páez, Awa Pit and Camsá are all located in the highlands, but show significant structural differences on the MDS maps. Especially the Páez and the Awa Pit languages are very interesting to analyze socioeconomically due to these calculated linguistic differences. Rappaport (1985) writes how the Páez people have maintained hegemony within their own areas, because the territories they inhabit have been determined by law. This has enabled the Páez to have their own distinct economic, social and political systems, so they do not necessarily have to seek socioeconomic advantages by contacting other groups or shifting to other languages. The Páez are mostly peasant cultivators, who identify farming as being one of the key characteristics of the Páez ethnic group. Maintaining this lifestyle in the highlands of the Andes in their own territories has definitely influenced the Páez language, because the result of the MDS measure shows how different Páez is compared to its neighbors Epena, Camsá and Awa Pit. Additionally, out of the nine languages used in this study, Páez has the highest number of speakers, 40,000, which reflects the hegemonic status of Páez within its own territories.

The people speaking the Barbacoan language Awa Pit, on the other hand, follow different socioeconomic patterns. The Awas also value land, since the Andes and its river valleys provide everything needed for living (Ordóñez 1992). In addition to farming and fishing, the Awas are also hunter-gatherers, which makes their livelihood quite versatile. Ordóñez mentions two interesting aspects of the social organization of the Awa group. First, the Awas have an egalitarian structure in their society, which means that work in the communities is divided equally between men and women, so women are not seen only as carers of the offspring. Second, the actual focal point of their social organization is the sibling group. A pair of siblings should always remain together, and they should preferably both marry their parallel cousins, which is a cousin from a parent's same-sex sibling. If this is not possible, "siblings move in pairs to form marital alliances

with people who are themselves kin" (Osborn 1968: 600, as cited in Ordóñez 1992). The peculiarity of this social organization might explain the structural differences of Awa Pit, if the people live and communicate only within their own group in the Andean highlands. This type of behavior does not support the acquisition of innovations, i.e. being influenced by other languages, so Awa Pit is not prone to change linguistically.

Attitudes towards language maintenance can definitely affect languages. An interesting difference is between Shuar and Jebero, which are Jivaroan and Cahuapanan languages spoken in the Andean Ecuador and in the Amazon lowlands of Peru, respectively. Their vitality status are quite different, because Shuar is a developing language and Jebero is already a shifting one. The reason for this difference can be explained by using socioeconomic factors. The Jebero people have centered around one village, and most of them are actually monolingual in Spanish, since the socioeconomic pressure of the prestige language has made people shift from their ethnic language to Spanish (Valenzuela 2010). Employment and education are two of the main reasons for this shift. Additional language shift occurs when the Jeberoans marry with non-Indians and gradually stop using their ethnic language.

The Shuar were semi-nomadic people, but eventually they also suffered from language shift due to missionaries, who taught them Spanish and encouraged the Shuars to live in settlements and not as nomads (Salazar 1977). In contrast to the Jebero people, Shuars wanted to prevent the language shift from continuing as it had by founding the Shuar Federation in 1964, which is a democratic organization administrating several social and economic affairs (Salazar 1977). The Federation has improved the Shuar identity of indigenous life style, which in turn has helped the vitality of the Shuar language. According to the 2007 census, Shuar had 35,000 speakers, and its vitality status is categorized as developing. So despite the language shift the Shuar managed to prevail by active language maintenance, which is not the case with Jebero. It is hard to say if these attitudes towards language maintenance have actually affected these two languages on a structural level, because Jebero and Shuar are located almost in the middle of the MDS map. Nevertheless, I believe that these socioeconomic factors have influenced these two languages on a wider scale, affecting their very existence.

These socioeconomic explanations have demonstrated how languages are not just abstract entities, but a part of our societies and identities. By looking at socioeconomic factors I have showed how they can help linguists in providing new information about the way languages function. If people feel like they would thrive and succeed in life by using a different language than their native one, the odds are that they will start using that other language. This has been the situation in South America for the indigenous people, since most people do not even really have a choice in the matter, if the hegemonic languages are used everywhere else in society, usually by people in power.

The social networks, i.e. the patterns of relationships people are involved in, also influence

indigenous people and their language use (Holmes 2013). Probably within most of the language groups studied in this thesis, the speakers use different languages with different people, according to their existing networks. Usually, if the network of relationships is an important one, it results in complying with the rest of the members, inevitably influencing the speakers' language. For example, when a person belonging to an ethnic group, such as the ones studied in this thesis, migrates from the rural areas to a major city, they might not have anyone to speak with in their native language and are forced to adapt to the city by learning a hegemonic language, which in South America means either Spanish or Portuguese. Therefore multilingualism is really common all around South America. Overall, languages are shifting and changing due to social, economic and demographic reasons, and the languages studied in this thesis are no exception.

6.1.2 Linguistic explanations

Considering how the language data consists of three language isolates and of languages from six different language families, the results of the MDS analysis should show obvious linguistic differences. Including a set of languages as diverse as possible was the whole idea of this study, because then I could calculate how different these languages actually are. The results showed that the structural differences are not as self-evident as they might have been. Some languages were not so different than others, and some showed rather significant differences, so in this section I will focus on linguistic factors and how they could explain the results.

The languages of South America, as well as of the research area, are highly discontinuous in their geographical distribution. Throughout the years languages have influenced one another through language contact, resulting in people changing their own linguistic tendencies. This process resonates with the result of the MDS analysis. When so many different languages are spoken in the same regions, they tend to have similar features through language contact (Csató et al. 2005). The contact is evidently a slow process, where "languages from several genetic groups that are located in the same geographical area will gradually come to share certain linguistic features" (Dixon & Aikhenvald 1999: 8). Linguistic features can be transferred from one language to another through borrowing, so that the borrowed features can be of any kind. For example, Aikhenvald (2007) lists four different ways how grammatical categories can be borrowed: the extensive and limited borrowing of the entire grammatical system, the borrowing of processes and the borrowing of syntactic constructions.

A detailed example of language contact is described by Aikhenvald (2006a). Aikhenvald specifies how Resígaro has been restructured almost completely due to the borrowing of linguistic features from unrelated Bora-Witotoan languages. This process has been quite significant, since Aikhenvald notes that a lexical comparison between Resígaro and Bora-Witotoan languages show how 24 % of the Resígaro lexicon are loan words. Additionally, the language contact has also in-

fluenced Resígaro's phonology, morphology and syntax. So to summarize, the excessive language contact has affected Resígaro considerably. Because of this restructuring, it is understandable why the MDS map shows significant differences between Resígaro and Yagua, despite their relative geographical proximity. As a language isolate, Yagua has not influenced Resígaro as much as other languages, so they have remained structurally different.

This complex occurrence of language contact shows how linguistic differences are not as apparent as they might have been thought to be. In general it is important to remember that language contact and linguistic borrowing do not magically alter languages overnight. It is a gradual process, and definitely a complex one. Without comparative-historical data on the languages used in this thesis, I can only reflect on the magnitude of language contact between these languages, as I have done in this section. Nevertheless, I believe these linguistic factors can explain why some of the languages are structurally similar and some structurally different than others.

6.2 Explaining the changes in linguistic diversity

The fluctuations of linguistic diversity were described in § 2.1.2.1 and § 2.1.2.2. In this section I will offer diachronic explanations for the changes in linguistic diversity in the tropical Andes and the Peruvian Amazon by using the punctuated equilibrium model. As this model suggests, the diversity in distinct areas has gone through stages of interruption and balance, i.e. punctuation and equilibrium. At some point in time the languages spoken near the Andes and in the Amazonian lowlands, and the groups of people speaking these languages have probably existed in a state of equilibrium. The punctuations in the area have since then caused populations to expand, which resulted in language splitting, thus creating more linguistic diversity. These punctuations have been caused by various reasons.

One massive punctuation was the development of agriculture. Due to unidentified reasons, some hunter-gatherers decided to cultivate land instead of foraging, which led to the early stages of agriculture in lowland Amazon (Dixon 1997). From the lowlands the agricultural way of living spread through South America, inevitably reaching the territories surrounding the Andes. The process of languages spreading into new territories is called linguistic expansion (Janhunen 2007). In general language spread can be defined as "an increase, over time, in the proportion of a communication network that adopts a given language or language variety for a given communicative function" (Cooper 1982: 6). It is important to underline that language spread does not mean that languages themselves are the ones who acquire speakers. It is the actual users who acquire languages and through them languages spread.

Linguistic expansion results in expansive and contractive languages. Expansive languages are those who manage to grow their geographical range through the expansion, whereas contractive languages are those who lose their territories, either partially or entirely (Janhunen 2007). The

development of agriculture caused a punctuation in population movement, leading to linguistic expansion, during which languages diversified when spreading into new territories. The area researched in this thesis is linguistically highly diverse, which supports the occurrence of linguistic expansion. When groups of people spread across South America in search of new lands for cultivation, their languages split into new ones. Additionally, the languages possibly spoken in the area before the arrival of other speech communities managed to survive this expansion.

The languages showcasing the highest amount of linguistic differences in the MDS map are the language isolates Camsá and Yagua. As language isolates, they are either proven not to have any genetic relations to other languages or they are the last remaining survivor of a language family otherwise gone extinct, as is the case with Yagua. Isolates are deemed as structurally different, and the MDS measures actually show just that. When other speech communities expanded into the areas surrounding the Andes, some might have ended up losing territories, and eventually speakers, but in the case of Camsá and Yagua, their geographical location might have helped them to survive the expansion. As mentioned in § 6.1.1.1, both Camsá and Yagua are languages existing in seclusion. Camsá is spoken in the high mountain ranges of the Andes, and Yagua is spoken in small remote communities across the Amazonian lowlands. Due to the their geographical isolation, these languages have managed to avoid the effect of the expansive languages, remaining as structurally different language isolates.

I believe that the effect of the agricultural punctuation on linguistic diversity was twofold. On the one hand, when people spread into new territories, their languages eventually split into new ones, creating linguistic diversity. On the other hand, the languages spoken in the area were not all absorbed by the expansive languages due to geographical isolation, so they remained as separate languages, reinforcing the linguistic diversity of the area. So in the end the punctuation caused by the development of agriculture enabled language diversification.

Another punctuation occurred when the Europeans expanded and migrated all across the world, eventually also into South America. This expansion was a devastating one, during which thousands of indigenous groups were either killed, enslaved or banished from their own territories. The most notorious aspect of the European invasion was the infectious diseases, which caused a punctuation in the linguistic diversity equilibrium. The epidemics spread throughout the continent, causing the majority of the indigenous people to perish. So when the agricultural punctuation enabled languages to diversify, the punctuation caused by the European invasion did quite the opposite. The invasion resulted in the loss of linguistic diversity, because entire indigenous societies were demolished.

However, when the Spanish invaded the areas nowadays known as Colombia, Ecuador and Peru, not all communities perished. Because the Andes offer remote and inaccessible territories where indigenous groups can live in seclusion, the Spanish occupation did not reach all the native peoples (Adelaar 2004). In its own rugged way, the Andes provided a place of refuge, which

enabled the languages spoken there also to survive. For example, the four Andean languages of Awa Pit, Camsá, Epena and Páez have all benefited from the seclusion the Andes have provided. Overall, the linguistic diversity in South America was severely affected by the Europeans. Yet, the physical geography of the area has maintained at least some part of the linguistic diversity.

The most recent punctuation in South America has been caused by the Industrial Revolution. Even though the Revolution started already in the 18th century, its effects are still felt in areas such as South America. The deforestation and the exploitation of the Amazon is not only a threat to the biodiversity of the area, but also to the linguistic diversity. Most of the indigenous groups living in the Amazonian rainforests have lost their entire habitats, so they have been forced to relocate somewhere else. This relocation can cause people to shift from one language to another, if they have to live in an area where other languages are more prominent. Eventually the language shift may result in the diminishing or even in the extinction of languages, which affects the overall linguistic diversity in South America.

Even though it is easy to pinpoint these few instances of punctuations, an important fact to keep in mind is that languages fluctuate all the time, even during an equilibrium period, meaning that they are not always entirely stable. Languages can unify and diversify depending on the prevailing circumstances. Some punctuations can increase linguistic diversity, while others can critically reduce it. Essentially several different factors can contribute to the changes in linguistic diversity, of which I explained three massive ones. These punctuations have generally affected the linguistic diversity in South America, but also the diversity in the tropical Andes and in the Peruvian Amazon.

6.3 Summary of the explanations

In this chapter I have offered possible explanations for the measured linguistic differences of nine indigenous languages spoken in the tropical Andes and in the Peruvian Amazon. I focused on using linguistic and non-linguistic factors as a base for my explanations, because languages are obviously independent entities, but they are connected to the environment and also to our identities as members of a society. The outcome of the MDS method can show underlying relationships between the measured objects. I believe it was important to explore these relationships, because they can show us something which would not be visible from just looking at tables of data.

This type of typological research is important in the field of linguistics, and explaining the reasons behind the results of a typological research is definitely worthwhile. With these explanations I suggested how linguistic, geographical and socioeconomic factors affect the structural differences of these languages. Linguistically, the contact-induced language change offered an approach to explain the differences. Language contact in the research area is visible for example in the borrowings from one language to another. Therefore some languages spoken in the same

regions show structural differences, when a language spoken in the area has borrowed linguistic elements from other languages. In general language contact is inevitable in South America, where hundreds of languages are spoken, of which most are distinctly discontinuous in their geographical distribution.

Geographically, I proposed a few interesting possibilities. First, it is evident that the Andes and the extensive river system of the Amazon offer natural barriers for indigenous peoples, which maintains their seclusion and prevents any kind of fusing into bigger languages from happening. An excellent example is the structural differences of four languages (Epena, Páez, Awa Pit, Camsá) spoken in the same region in the Andean Colombia. Despite the close proximity they showcase visible structural differences in the MDS map, which can be explained by the effect the geographical factors have on the area. The Andes provide everything these societies need, from isolation to nourishment. Second, the rivers have assisted the seclusion of some languages in providing remote locations for villages. This has resulted in quite distinct structural differences compared to the other languages. On the other hand, the extensive river system might explain why some languages were quite similar on the MDS map despite their geographical distances, since the rivers can provide means of transportation, enabling possible contact between different ethnic groups.

Socioeconomically, the research area showed lots of interesting aspects for explanations. For the most part, the languages of this area can be categorized as shifting languages, since the socioeconomic factors have caused indigenous people to abandon their native language in favor of a prestige language. If people want to get educated or want to be employed, the only possible solution for them has been to switch from one language to another. Some people do not even resent the situation, because they just cannot see how their native indigenous language could benefit them as members of a society. This shift has been especially rapid among young people, and their mobility has affected several languages and caused them to lose speakers.

Another accelerating factor of language shift is the intermarriage between groups. If some groups only allow intragroup marriages, their language might have persevered longer, resulting in structural differences. These types of cases are also apparent in the languages studied in this thesis. From the nine languages used in the MDS analysis only two have actually shown how some socioeconomic factors can actually help maintain the speakers from shifting languages. Páez has maintained hegemony due to territories determined by law, enabling distinct economic, social and political systems, while the Shuar have founded its own federation which administers several social and economic affairs. This type of active language maintenance can help ethnic groups to maintain their identities, and therefore to keep their languages as structurally diverse as possible. Language shift will be inevitable for groups without active maintenance.

Based on the explanations it is evident that the relationship between language and its environment is really complex. I believe that these linguistic and non-linguistic explanations have

strengthened the perception of this relationship. There is also an intertwined relationship between the linguistic and non-linguistic factors. For example, when the colonizers invaded the Yagua region in order to harvest rubber trees, the indigenous people fled. The extensive river system offered the best way to transport the groups further into the jungle, making them even more secluded from other groups than before. The economic invasion led to the migration of those people, which was intensified by the topography of the region, resulting in ethnic groups living in isolation, ensuring that their language would live in isolation as well. This is the perfect example why I can verify my hypothesis by stating that there is a correlation between the ecological environment of languages and their linguistic differences.

By conducting a statistical analysis I was able to show how linguistic differences can be measured. I used the MDS method, which resulted in the visual representation of the differences. Additionally, I wanted to discover if language isolates could be structurally different from languages which have existing sister languages in their language families. The results showed significant differences between two isolates and the other languages. This could suggest that linguistic differences can be affected by the languages' existing genealogical relationships.

I also wanted to find out if an ecological approach and non-linguistic parameters can actually be used to explain linguistic differences. I believe the explanations presented in this chapter show how both geographical and socioeconomic explanations can be used to explain structural differences. In some cases these explanations can even shed light on just how much factors such as geographical location can affect the entire existence of languages. Overall, the non-linguistic explanations have underlined the complex relationship between language and its environments. So to conclude, the explanations offered in this chapter justify the use of non-linguistic parameters in explaining linguistic differences.

In addition to explaining linguistic differences, I was able to clarify why and how the linguistic diversity of the area has changed. As Adelaar (2004: 4) has noted, there has been "an alternation between periods of greater communication and integration of different peoples and languages, and periods of fragmentation and individual development". This quote summarizes the fluctuations in the diversity quite well. The area has been affected by different punctuations, which have either diversified the languages or diminished the linguistic diversity. Whether the area is in a state of equilibrium or suffering from a punctuation, it is clear that in general languages are always changing.

7 Conclusion

In this thesis I explored the linguistic differences of nine indigenous languages spoken in the tropical Andes and in the Peruvian Amazon by using a statistical method called multidimensional scaling. Additionally, I offered linguistic and non-linguistic explanations for these differences. I also focused on the linguistic diversity of the area by explaining diachronically how it has fluctuated. The analysis showed how new methods of studying languages and their differences can be used to understand the processes shaping languages even better. Based on the results I was able to answer the research questions, one of which led to interesting findings on language isolates. The findings suggest that the significant structural differences of language isolates compared to the other languages in the data set could be due to their status as the only living member of a specific language family. This suggestion strengthens the position of language isolates as languages which have inspired and should continue to inspire additional research. As a result of the conducted research I was able to verify my hypothesis and conclude that there is a correlation between the ecological environment of languages and their linguistic differences.

This thesis is an interdisciplinary case study of a specific area and its languages, which combines quantitative and qualitative approaches. Therefore the results and their interpretations are only meant to be considered in relation to the research area and the languages studied in this thesis. However, the methodology used in this study can be applied to other areas of high linguistic diversity and also to other languages. The problems concerning the study lie in the breadth of the data, which is quite restricted. Due to the discrepancies in the WALS data, the language data only consists of languages which have the exact same features represented in the feature data. This resulted in the data consisting of nine languages, which share only six features. With a better data set linguists could study linguistic differences and linguistic diversity in greater depth, which might results in astonishing new discoveries on the relationship between languages and the environment.

Overall, this type of case study is a starting point for future research and it can be developed in many different directions. In general, the study could be replicated by adding more languages and linguistic parameters, which could lead to more typologically representative results. Other statistical methods might be valuable to use either together with or instead of multidimensional

scaling. As a statistical method, multidimensional scaling could be used to study the linguistic differences in an entire country or even in a continent, or it could be used to study the internal differences of a specific language family. Additionally, language isolates and their structural differences could be studied from a wider point of view, which might help linguists in understanding the otherwise extinct language families more effectively. The explanations for linguistic differences and linguistic diversity could be expanded to include new and more detailed explanations, such as cultural and archaeological ones. Additional explanations would generate even more interdisciplinary studies, which could really benefit the field of linguistics.

As it is clear, this thesis is definitely a starting point for further studies, which is emphasized by the sheer number of possible applications. Because languages are always changing, there will always be a need for studies in linguistic diversity and in linguistic differences.

References

Adelaar, Willem F. H. 2004. *The languages of the Andes*. Cambridge: Cambridge University Press.

Adelaar, Willem F. H. 2012. Languages of the Middle Andes in areal-typological perspective: Emphasis on Quechuan and Aymaran. In Campbell, Lyle & Grondona, Veró (eds.), *The indigenous languages of South America: A comprehensive guide*, 575–624. Berlin: Walter de Gruyter.

Aikhenvald, Alexandra Y. 2006a. Areal diffusion, genetic inheritance, and problems of subgrouping: A North Arawak case study. In Aikhenvald, Alexandra Y. & Dixon, R. M. W. (eds.), *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, 167–194. Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. 2006b. Grammars in contact: A cross-linguistic perspective. In Aikhenvald, Alexandra Y. & Dixon, R. M. W. (eds.), *Grammars in contact: A cross-linguistic typology*, 1–66. Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. 2007. *Language contact in Amazonia*. Oxford: Oxford University Press.

Axelsen, Jacob Bock & Manrubia, Susanna. 2014. River density and landscape roughness are universal determinants of linguistic diversity. *Proceedings of the Royal Society of London B: Biological Sciences* 281(1784). `http://dx.doi.org/10.1098/rspb.2013.3029`. [Accessed 1 March 2018].

Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Blench, Roger. 2008. Accounting for the diversity of Amerindian languages: Modelling the settlement of the New World. *Archaeology Research Seminar, RSPAS.* `http://www.rogerblench.info/Archaeology/New%20World/Peopling%20of%20the%20New%20World%20Canberra%20paper.pdf`. [Accessed 27 December 2017].

Blouet, Brian W. & Blouet, Olwyn M. 2015. *Latin America and the Caribbean: A systematic and regional survey*. 7th edn. New Jersey: John Wiley & Sons.

Borg, Ingwer & Groenen, Patrick J. F. 2005. *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Borin, Lars. 2013. The why and how of measuring linguistic differences. In Borin, Lars & Saxena, Anju (eds.), *Approaches to measuring linguistic differences*, 3–25. Berlin: De Gruyter Mouton.

Borin, Lars & Saxena, Anju (eds.). 2013. *Approaches to measuring linguistic differences*. Berlin: De Gruyter Mouton.

Bowens, Amanda (ed.). 2011. *Underwater archaeology: The NAS guide to principles and practice*. New Jersey: John Wiley & Sons.

Campbell, Lyle. 2012. Classification of the indigenous languages of South America. In Campbell, Lyle & Grondona, Veró (eds.), *The indigenous languages of South America: A comprehensive guide*, 59–166. Berlin: Walter de Gruyter.

Chambers, Jack K. & Trudgill, Peter. 2004. *Dialectology*. Cambridge: Cambridge University Press.

Chiswick, Barry R. & Miller, Paul W. 2005. Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development* 26(1). 1–11. `http://ftp.iza.org/dp1246.pdf`. [Accessed 22 February 2018].

Clarke, Robert K. 1993. Non-parametric multivariate analyses of changes in community structure. *Austral Ecology* 18(1). 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x. [Accessed 20 February 2018].

Cooper, Robert L. (ed.). 1982. *Language spread: Studies in diffusion and social change*. Bloomington: Indiana University Press.

Croft, William & Poole, Keith T. 2008. Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics* 34. 1–37. https://legacy.voteview.com/pdf/Mds_paper.pdf. [Accessed 18 April 2018].

Csató, Éva Ágnes, Isaksson, Bo & Jahani, Carina. 2005. *Linguistic convergence and areal diffusion: Case studies from Iranian, Semitic and Turkic*. London: Psychology Press.

de Busser, Rik. 2015. The influence of social, cultural, and natural factors on language structure. In de Busser, Rik & LaPolla, Randy J. (eds.), *Language structure and environment*, 1–28. Amsterdam: John Benjamins.

Dixon, R. M. W. 1997. *The rise and fall of languages*. Cambridge: Cambridge University Press.

Dixon, R. M. W. & Aikhenvald, Alexandra Y. 1999. Introduction. In Dixon, R. M. W. & Aikhenvald, Alexandra Y. (eds.), *The Amazonian languages*, 1–21. Cambridge: Cambridge University Press.

Dryer, Matthew S. 2013a. Coding of nominal plurality. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The World Atlas of Language Structures online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/33.

Dryer, Matthew S. 2013b. Negative morphemes. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The World Atlas of Language Structures online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/112.

Dryer, Matthew S. 2013c. Order of subject, object and verb. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The World Atlas of Language Structures online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/81.

Dryer, Matthew S. 2013d. Position of tense-aspect affixes. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The World Atlas of Language Structures online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/69.

Dryer, Matthew S. 2013e. Prefixing vs. suffixing in inflectional morphology. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The World Atlas of Language Structures online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/26.

Dryer, Matthew S. & Haspelmath, Martin (eds.). 2013. *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Dunn, Terry. 2003. Resource partitioning: Definition, theory & examples https://study.com/academy/lesson/resource-partitioning-definition-theory-examples.html. [Accessed 23 October 2017].

ECLAC. 2014. Guaranteeing indigenous people's rights in Latin America: Progress in the past decade and remaining challenges http://www.cepal.org/publicaciones/default.asp?idioma=IN. [Accessed 5 April 2018].

Edwards, John. 2012. *Multilingualism: Understanding linguistic diversity*. London: Bloomsbury Publishing.

Eliasson, Stig. 2015. The birth of language ecology: Interdisciplinary influences in Einar Haugen's "The ecology of language". *Language Sciences* 50. 78–92. https://doi.org/10.1016/j.langsci.2015.03.007. [Accessed 26 April 2018].

Everitt, Brian & Hothorn, Torsten. 2006. *An introduction to applied multivariate analysis with R*. New York, NY: Springer.

Fill, Alwin. 2007. Language contact, culture and ecology. In Hellinger, Marlis & Pauwels, Anne

(eds.), *Handbook of language and communication: Diversity and change*, 177–207. Berlin: Mouton de Gruyter.

Gavin, Michael C., Botero, Carlos A., Bowern, Claire, Colwell, Robert K., Dunn, Michael, Dunn, Robert R., Gray, Russell D., Kirby, Kathryn R., McCarter, Joe, Powell, Adam, Rangel, Thiago F., Stepp, John R., Trautwein, Michelle, Verdolin, Jennifer L. & Yanega, Gregor. 2013. Towards a mechanistic understanding of linguistic diversity. *BioScience* 63(7). 524–535. http://dx.doi.org/10.1525/bio.2013.63.7.6. [Accessed 7 September 2017].

Gavin, Michael C. & Stepp, John R. 2014. Rapoport's Rule revisited: Geographical distribution of human languages. *PLOS ONE* 9(9). 1–8. https://doi.org/10.1371/journal.pone.0107623. [Accessed 28 December 2017].

Goddard, Ives & Campbell, Lyle. 1994. The history and classification of American Indian languages: What are the implications for the peopling of the Americas. *Method and Theory for Investigating the Peopling of the Americas. Oregon State University, Corvallis: Center for the Study of the First Americans* 189–207.

Gorenflo, Larry J., Romaine, Suzanne, Mittermeier, Russell A. & Walker-Painemilla, Kristen. 2012. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *PNAS* 109(21). 8032–8037. https://doi.org/10.1073/pnas.1117511109. [Accessed 10 November 2017].

Gower, John C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27(4). 857–871. http://links.jstor.org/sici?sici=0006-341X%28197112%2927%3A4%3C857%3AAGCOSA%3E2.0.CO%3B2-3. [Accessed 16 April 2018].

Gries, Stefan. 2013. *Statistics for linguistics with R: A practical introduction*. Berlin: Walter de Gruyter.

Haugen, Einar. 2001. The ecology of language. In Fill, Alwin & Mühlhäusler, Peter (eds.), *The ecolinguistics reader: Language, ecology and environment*, 57–66. London: Continuum.

Heine, Bernd & Kuteva, Tania. 2005. *Language contact and grammatical change*. Cambridge: Cambridge University Press.

Holmes, Janet. 2013. *An introduction to sociolinguistics*. London: Routledge.

Hout, Michael C., Papesh, Megan H. & Goldinger, Stephen D. 2013. Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science* 4(1). 93–103. https://doi.org/10.1002/wcs.1203. [Accessed 15 February 2018].

Janhunen, Juha A. 2007. Typological expansion in the Ural-Altaic belt. *Incontri Linguistici* 30. 71–83.

Johnson, Keith. 2008. *Quantitative methods in linguistics*. New Jersey: Blackwell Publishing.

Kerr, Jeremy T. & Packer, Laurence. 1997. Habitat heterogeneity as a determinant of mammal species richness in high-energy regions. *Nature* 385(6613). 252–254. http://dx.doi.org/10.1038/385252a0. [Accessed 28 December 2017].

Lucas, Christopher. 2015. Contact-induced language change. In Bowern, Claire & Evans, Bethwyn (eds.), *The Routledge handbook of historical linguistics*, 519–536. London: Routledge.

Lupyan, Gary & Dale, Rick. 2015. The role of adaptation in understanding linguistic diversity. In de Busser, Rik & LaPolla, Randy J. (eds.), *Language structure and environment*, 289–316. Amsterdam: John Benjamins.

Lynch, Thomas F. 1999. The earliest South American lifeways. In Salomon, Frank & Schwartz, Stuart B. (eds.), *The Cambridge history of the Native Peoples of the Americas*, vol. 3, 188–263. Cambridge: Cambridge University Press.

Mace, Ruth & Pagel, Mark. 1995. A latitudinal gradient in the density of human languages in North America. *Proceedings of the Royal Society of London B* 261(1360). 117–121. http://rspb.royalsocietypublishing.org/content/261/1360/117. [Accessed 15 September 2017].

Maddieson, Ian. 2013. Consonant-vowel ratio. In Dryer, Matthew S. & Haspelmath, Martin (eds.),

The World Atlas of Language Structures online, Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/3.

Maechler, Martin, Rousseeuw, Peter, Struyf, Anja, Hubert, Mia & Hornik, Kurt. 2017. *cluster: Cluster analysis basics and extensions*. R package version 2.0.6.

McEnery, Anthony & Wilson, Anita. 2001. *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Moore, Joslin L., Manne, Lisa, Brooks, Thomas, Burgess, Neil D., Davies, Robert, Rahbek, Carsten, Williams, Paul & Balmford, Andrew. 2002. The distribution of cultural and biological diversity in Africa. *Proceedings of the Royal Society of London B: Biological Sciences* 269(1501). 1645–1653. http://rspb.royalsocietypublishing.org/content/269/1501/1645. [Accessed 15 September 2017].

Mulligan, Connie J. & Szathmáry, Emöke J. E. 2017. The peopling of the Americas and the origin of the Beringian occupation model. *American Journal of Physical Anthropology* 162. 403–408. https://doi.org/10.1002/ajpa.23152. [Accessed 27 December 2017].

Muysken, Pieter. 2012. Contacts between indigenous languages in South America. In Campbell, Lyle & Grondona, Veró (eds.), *The indigenous languages of South America: A comprehensive guide*, 235–258. Berlin: Walter de Gruyter.

Nerbonne, John. 2003. Linguistic variation and computation (invited talk). In *10th conference of the european chapter of the association for computational linguistics*, 3–10. http://www.aclweb.org/anthology/E03-1088. [Accessed 20 February 2018].

Nerbonne, John & Hinrichs, Erhard. 2006. Linguistic distances. In *Proceedings of the workshop on linguistic distances*, 1–6. Association for Computational Linguistics. http://www.aclweb.org/anthology/W06-1101. [Accessed 20 February 2018].

Nettle, Daniel. 1998. Explaining patterns of language diversity. *Journal of Anthropological Archaeology* 17. 354–374. https://doi.org/10.1006/jaar.1998.0328. [Accessed 5 September 2017].

Nettle, Daniel. 1999. *Linguistic diversity*. Oxford: Oxford University Press.

Nettle, Daniel. 2009. Ecological influences on human behavioural diversity: A review of recent findings. *Trends in Ecology & Evolution* 24(11). 618–624. https://doi.org/10.1016/j.tree.2009.05.013. [Accessed 20 September 2017].

Nettle, Daniel & Romaine, Suzanne. 2000. *Vanishing voices*. Oxford: Oxford University Press.

Nichols, Johanna. 1990. Linguistic diversity and the first settlement of the New World. *Language* 475–521. http://www.jstor.org/stable/414609. [Accessed 27 December 2017].

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: The University of Chicago Press.

Oksanen, Jari. 2003. R: Opas ekologeille [a guide for ecologists] http://cc.oulu.fi/~jarioksa/opetus/rekola/Rekola.pdf. [Accessed 24 January 2018].

Ordóñez, Pedro Vicente Obando. 1992. *Awa-Kwaiker: An outline grammar of a Colombian/Ecuadorian language, with a cultural sketch*. Ann Arbor, MI: University Microfilms.

Payne, Doris L. 1985. *Aspects of the grammar of Yagua: A typological perspective*: Los Angeles, CA: University of California dissertation.

Peterson, Barbara Bennett. 2011. *Peopling of the Americas: Currents, canoes, and DNA*. New York: Nova Science Publishers.

Population Reference Bureau. 2017. World population data sheet https://www.prb.org/2017-world-population-data-sheet/. [Accessed 6 April 2018].

R Core Team. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Salazar, Ernesto. 1977. *An Indian federation in lowland Ecuador*. Copenhagen: IWGIA.

Schepens, Job, Van der Slik, Frans & Van Hout, Roeland. 2013. The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language. In Borin, Lars & Saxena, Anju (eds.), *Approaches to measuring linguistic differences*, 199–229. Berlin: De Gruyter Mouton.

Schmidtke-Bode, Karsten & Hetterle, Katja. 2008. Multivariate analysis of linguistic data: Multidimensional scaling `http://www.kschmidtkebode.de/Multidimensional%20scaling_July%204%202008.pdf`. [Accessed 15 February 2018].

Simons, Gary F. & Fennig, Charles D. (eds.). 2018. *Ethnologue: Languages of the World*. Dallas, TX: SIL International. `http://www.ethnologue.com`.

StatSoft, Inc. 2013. Electronic statistics textbook. `http://www.statsoft.com/textbook/`.

Thomason, Sarah G. & Kaufman, Terrence. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley, CA: University of California Press.

Valenzuela, Pilar M. 2010. Ethnic-racial reclassification and language revitalization among the Shiwilu from Peruvian Amazonia. *International Journal of the Sociology of Language* 202. 117–130. `https://doi.org/10.1515/ijsl.2010.017`. [Accessed 4 March 2018].

Veblen, Thomas T., Young, Kenneth R. & Orme, Antony R. 2015. *The physical geography of South America*. Oxford: Oxford University Press.

Venables, William N. & Ripley, Brian. 2002. *Modern applied statistics with S*. 4th edn. New York: Springer. `http://www.stats.ox.ac.uk/pub/MASS4`.

Venables, William N. & Smith, David M. 2009. *An introduction to R*. Bristol: Network Theory Ltd.

Wendel, John N. 2005. Notes on the ecology of language. *Bunkyo Gakuin University Academic Journal* 5. 51–76. `https://www.u-bunkyo.ac.jp/center/library/image/fsell2005_51-76.pdf`. [Accessed 20 October 2017].

Winford, Donald. 2005. Contact-induced changes: Classification and processes. *Diachronica* 22(2). 373–427. `https://doi.org/10.1075/dia.22.2.05win`. [Accessed 3 April 2018].